CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

D3.2

Series of five short survey papers on methodological concerns

Authors: Christof Schöch, Joanna Byszuk, Julia Dudar, Maciej Eder, Evegniia Fileva, Andressa Gomide, Lisanne van Rossum, Artjoms Šeļa, Karina van Dalen-Oskam

Date: March 31, 2023

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Project Acronym: CLS INFRA
Project Full Title: Computational Literary Studies Infrastructure
Grant Agreement No.: 101004984

Deliverable/Document Information

Deliverable No.: D3.2
Deliverable Title: Series of five short survey papers on methodological concerns
Authors: Christof Schöch, Joanna Byszuk, Julia Dudar, Maciej Eder, Evegniia Fileva, Andressa Gomide, Lisanne van Rossum, Artjoms Sela, Karina van Dalen-Oskam
Dissemination Level: PUBLIC

Document History

Version/Date	Changes/Approval	Author/Approved by
March 31, 2023	v1.0.0: complete version	Christof Schöch and all authors
February 28, 2023	v0.1.0: initial draft version	Christof Schöch and all authors

Survey of Methods in Computational Literary Studies

=CLS INFRA D3.2: Series of five short survey papers on methodological issues

Christof Schöch, Julia Dudar, Evegniia Fileva (editors)

2023-05-05

Table of contents

Front Matter

Abstract

The aim of this publication is to document and describe current, widespread research practices in Computational Literary Studies (CLS), based on a large collection of publications that have been published in this field over the last approximately ten years. The perspective of this survey is primarily descriptive: it aims to document current, widespread practices as the authors were able to observe them in the published literature. In this sense, the survey, while far from exhaustive, can also serve as an annotated bibliography of sorts and as a guide to further reading. Despite the fact that this survey is not intended as an introductory textbook, it can nevertheless also serve as an introduction to several research areas and issues that are prominent within CLS as well as to several key methodological concerns that are of importance when performing research in CLS.

Publication formats

Please note that this work is distributed in several formats, a static PDF (for reference), an HTML version (for flexible reading) and a full set of all production files (for documentation):

- The HTML version can be found on the CLS INFRA website at methods.clsinfra.io.
- The PDF version can be found on Zenodo in the Computational Literary Studies Infrastructure community under the following DOI: 10.5281/zenodo.7892112.
- All files needed to render the publication (in a ZIP archive), including a howto-file describing the process, is also included in the Zenodo repository.

Funding acknowledgement

This survey of research practices in Computational Literary Studies was prepared by a group of contributors in the framework of the Starting Community *Computational Literary Studies Infrastructure* (CLS INFRA) funded by the European Commission in the Horizon 2020 programme under Grant agreement ID 101004984.

Citation suggestion

Christof Schöch, Julia Dudar, Evegniia Fileva, eds. (2023). *Survey of Methods in Computational Literary Studies* (= CLS INFRA D3.2: Series of Five Short Survey Papers on Methodological Issues). With contributions by Joanna Byszuk, Julia Dudar, Evegniia Fileva, Andressa Gomide, Lisanne van Rossum, Christof Schöch, Artjoms Šeļa and Karina van Dalen-Oskam. Version 1.1.0, May 5, 2023. Trier: CLS INFRA. URL: https://methods.clsinfra.io, DOI: 10.5281/zenodo.7892112.

Please note that this is the citation suggestion for the entire survey. If you use or cite specific sections of the survey, please use the relevant citation suggestion instead.

```
@book{schoch_2023_survey,
  title = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/},
  editora = {Byszuk, Joanna and Dudar, Julia and Fileva, Evgeniia
  and Gomide, Andressa and van Rossum, Lisanne and Schöch, Christof
  and Šeļa, Artjoms and van Dalen-Oskam, Karina},
  editoratype = {collaborator},
  langid = {english}
}
```

General Introduction

Christof Schöch (Trier)

Welcome to this methodological survey of Computational Literary Studies (CLS). The aim of this publication is to document and describe current, widespread research practices in CLS, based on a large collection of publications that have been published in this field over the last approximately ten years. The perspective of this survey is primarily descriptive: it aims to document current, widespread practices as the authors were able to observe them in the published literature. In this sense, the survey can also serve as an annotated bibliography of sorts and as a guide to further reading. Despite the fact that this survey is not intended as an introductory textbook, it can nevertheless also serve as an introduction to several research areas or issues that are prominent within CLS as well as to several key methodological concerns that are of importance when performing research in CLS.

Areas of research and steps in the research process

The research areas or issues covered by this survey are the following, and correspond to significant strands of CLS research:

- Stylometric authorship attribution, that is the identification of possible authors of a text of anonymous or disputed authorship using quantitative indicators (for an introduction, see Chapter 5);
- Literary genre analysis, that is the analysis of the similarities and differences between groups of texts belonging to related but distinct literary genres or subgenres (see Chapter 10);
- Literary history, that is the investigation of patterns and developments that are determined by factors such as literary periods and movements or the development of literature over time (see Chapter 15);
- Analysis of gender, that is the investigation of patterns, differentiations and other phenomena related either to the biological sex or social gender of authors or to the ascribed sex and gender identities of fictional characters in literary works (see Chapter 20);
- Finally, the analysis of canonicity and prestige and the ways these attributions to authors and texts are related both to textual and extra-textual factors and processes (see Chapter 25).

The key methodological concerns covered by this survey, in turn, are the following, and are directly related to a certain number of key steps in the research process:

- Corpus building, that is the design and composition of corpora in such a way that they best support the investigation of one or several research questions (for an introduction, see Chapter 1);
- Preprocessing and annotation, that is the process of preparing texts selected for inclusion into a corpus, or belonging to a previously designed corpus, by way of text encoding, data cleaning, token-level annotation and document-level metadata collection (see Chapter 2);

- Analysis, that is the performance of qualitative or (in particular) quantitative methods of operationalizing or formalizing specific literary phenomena and the investigation of the nature, prevalence, and distribution of such phenomena in literary corpora (see Chapter 3);
- Finally, the formal evaluation of the robustness, generalizability, explainability, performance and/or significance of the analyses performed in the previous step (see Chapter 4).

Some practical matters

The two perspectives described above, the research areas or issues, on the one hand, and the methodological concerns or steps in the research process, on the other, structure this survey, which is presented as a two-dimensional grid of short texts. As a consequence, this survey can be read in at least three manners: Readers who are primarily interested in different aspects of a given research issue, such as authorship attribution or genre analysis, may want to read all the texts in the relevant column, from top to bottom. Readers, however, who are rather interested in different aspects of a given step in the research process or methodological concerns, such as corpus building or evaluation, you may want to read all the texts in the relevant row, from left to right. A set of short introductions to each research issue and each methodological concern provide orientation to readers adopting either approach. Finally, readers are of course welcome as well to dive right in and read texts in any order they prefer.

Please note that bibliographic references are included in each text using the 'author date'-system in brakets within the text. The full references for a given section, including both works cited as well as additional references recommended for further reading, can be obtained by clicking on the link provided at the end of each individual chapter. A list of all cited references is also provided for convenience's sake. In addition, all cited references as well as further readings are available in the CLS Bibliography that also documents the corpus of publications that is the foundation of this survey.

Citation suggestion

Christof Schöch (2023): "General Introduction". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/general-intro.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{schoch_2023_general,
  title = {General Introduction},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Schöch, Christof},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/general-intro.html},
  langid = {english}
}
```

Part I

Introductions to the research steps

This part of the survey is devoted to general, introductory remarks about the individual steps in the research process covered by this methodological survey of Computational Literary Studies.

1 Introduction to Corpus Building

Evgeniia Fileva (Trier)

1.1 What is corpus building?

Not only are corpora and text collections a way of preserving literary texts for the long-term transmission of literary heritage, but they are also an essential foundation for research in both Digital Humanities generally and Computational Literary Studies in particular. The number of available corpora is large, and increases every year. This is primarily due to the fact that most research projects have specific requirements with respect to the corpus of texts being used. As a consequence, many corpora are created for specific research purposes and are therefore as diverse as the projects they support.

Corpus building is used here as a cover term for a range of activities related to the creation of both linguistic corpora specifically – i.e. systematically composed and linguistically-annotated sets of machine-readable texts (McEnery and Hardie 2012) – and text collections more generally – i.e. small or large collections of literary texts or historical documents with various degrees of rigour in their composition and a broad range of possible annotations. These activities first of all include conceptual work, such as deciding on the scope and size of a corpus, designing the criteria for the composition of the corpus, designing the text encoding scheme and modeling the information captured by the metadata. But they also include the practical implementation of these design decisions, such as identifying texts for inclusion in the corpus, digitizing the texts (if they are not available in digital form), applying the encoding scheme to the texts, collecting the required metadata and making the corpus available to others.

All in all, our analysis of the research literature in CLS covered here contains relatively few studies specifically devoted to the issue of corpus building. It appears that knowledge on this topic is mostly derived from adjacent, relevant fields, such as Corpus Linguistics, text encoding and scholarly digital editing, from which many lessons are probably taken. A limited number of studies, however, do exist also within CLS proper that describe and reflect on the corpus building process in particular cases and for particular purposes. These are presented in other chapters of this survey that treat corpus building for specific areas of investigation.

Some publications from within CLS also reflect in a more general or theoretical manner on what corpora are, how they can be designed and what kind of role they play in the research-based process. This category includes much work by Katherine Bode, concerned in particular with the biases baked into corpora due to the ways in which they have been created, often using material that has become available due to many historical and contextual factors rather than only scholarly concerns (e.g. Bode 2020). A recent addition to this body of work is Michael Gavin's book on *Literary Mathematics* (Gavin 2023), in which the author underlines the importance of the construction of the corpus as a place where words (in the texts themselves) and contexts (through the metadata describing each individual text) come together and support multiple avenues of mathematical analysis. A more modest summary of best practices in the design of datasets, including corpora for research in CLS, has been proposed by Christof Schöch (Schöch 2017a).

1.2 What are the major issues in corpus building?

1.2.1 Corpus size

The size of a corpus can vary considerably. In addition, the size of a corpus may be calculated in various ways, notably as the number of texts it contains (irrespective of the texts' length), or as the number of tokens (words) or types (distinct words) it contains. In terms of the number of texts, corpora can range from a relatively small size, as in the *European Literary Text Collection* (ELTeC), where each separate corpus contains 100 novels, to millions of books, as in the case of the *HathiTrust Collection*, made up of over 17 million items (Schöch et al. 2021; Organisciak et al. 2014). But keep in mind that novels are relatively long texts, in terms of the number of tokens, so that virtually all ELTeC corpora of just 100 novels contain more than 5 million tokens each, whereas the *Diachronic Spanish Sonnet Corpus* (DISCO), with more than 4300 texts, is made up of slightly more than 400.000 tokens (Ruiz Fabo, Martínez Cantón, and Calvo Tello 2018).

The estimated corpus size is generally determined according to the goal of the researchers building the corpus. If the goal is to show the linguistic and/or literary diversity of a language or literary tradition, e.g. to preserve literary heritage in a digital format, or to provide a background for further research, a corpus can consist of a large number of texts. However, a small, representative corpus can also serve as a useful corpus for DH and CLS research. In addition, one cannot exclude technical limitations, which do not always allow working with a large amount of data, and if this is the case, scholars are limited to a smaller amount of texts. Overall, corpus size depends on the goals of a study, and larger databases may be necessary to capture either rarer or more diverse phenomena or to discover and test for trends over time (Schöch 2017a).

1.2.2 Markup, annotation, format

Markup is a way to add information to the corpus in a machine-readable format. Markup can be divided into two types: document-level markup and annotations of parts of the texts (Reppen 2022). Depending on the community of origin, corpora may be encoded in one of a number of widespread formats. The simplest way to provide the texts in a corpus is the plain text format. However, in this case, neither information on textual structure nor token-based linguistic annotation can be provided. In addition, metadata can in this case only be transmitted in a separate file or via the filenames. If filenames are used for this purpose, the names of units in the corpus should be standardized and reduced to the same format, containing, for example, the name of the book, author, language, etc. A more suitable method is to use filenames as identifiers, but store metadata in a separate, tabular file for the corpus as a whole using the identifiers to link files and relevant metadata entries.

In DH and CLS, the accepted standard for textual encoding are the Guidelines of the *Text Encoding Initiative* (TEI). This XML-based format allows for the inclusion of detailed metadata, encoding of textual structure as well as token-based linguistic annotation (see (Burnard 2014)). Using a formal language like XML to represent analytic annotation in a text has a significant advantage in automatic validation. This means that the annotation used in a document can be checked to see if it conforms to a previously defined model of which types of annotation are permitted and in which contexts (Burnard 2014). Examples of recent literary corpora that use XML-TEI include the *Deutsches Textarchiv*, the *Diachronic Spanish Sonnet Corpus*, the *European Literary Text Collection*, *DraCor*, *TextGrid's Digitale Bibliothek*, *Théâtre classique* and many more.

1.2.3 Metadata

Metadata contains information about various aspects of the books included in the corpus, such as about the author, publication, format, etc. Metadata is usually contained in the XML/TEI markup of the document. This kind of markup is commonly occurring in the Deutsches Textarchiv, CoNSAA, and ELTeC corpora. Metadata, i.e. knowledge about the text, can also be stored as a CSV table. This type of metadata storage can be seen, for instance, in the Zeta project corpus. The table, published in the Github repository, contains information about each novel, including genre and subgenre. Project Gutenberg has a feature in its editorial metadata such as the absence of print source publication date information. Since the corpus includes e-books, this criterion has been replaced by a release date. In addition, the Gutenberg project uses XML/RDF format for metadata. However, this scheme has not always been used in the project; previously metadata was stored in the MARC (Machine-Readable Cataloging) format. This format is now used by the HathiTrust corpus, which stores content and bibliographic information. You can read more about the metadata requirements of the HathiTrust project and the process of data submission on their website.

Corpora can provide different kinds of document-level information about texts (metadata). There are various approaches to classifying metadata types. The standard metadata types defined by NISO include descriptive, structural, and administrative kinds (see National Information Standards Organization). In some studies, such as those by Burnard and Calvo Tello, there are slight differences from the NISO typology. For example, Calvo Tello lists editorial metadata instead of structural metadata (Calvo Tello 2021). Burnard includes all 4 types of metadata (Burnard 2014). The ELTeC corpus also distinguishes 4 types of metadata, similar in structure to Burnard's typology. Storing metadata in the markup according to the standardized TEI format, where the TEI Header provides a large number of metadata elements and attributes at the beginning of each file is a very flexible and powerful solution.

1.2.4 Data accessibility and copyright

orpora and corresponding corpus documentation (information about metadata, format, DOI and URI identification numbers for each text, etc.) can then be placed in appropriate repositories, such as for example libraries and archives, repositories like Zenodo or Figshare or cross-project infrastructure initiatives (e.g. DARIAH or CLARIN). In this way, the data are published, archived and can be used, verified or reproduced by other researches for further study. It is important to make sure that the data meet standards (e.g. TEI encoding) and are technically available (Schöch 2017a).

It's not always the case that the materials from the corpus are fully open access or in the public domain. If corpus consists of closed-access materials, permission must be obtained for their use as well as for the use of the corpus itself. In such scenarios, the corpus cannot be made available to third parties. However, there are solutions and proposals available that support so-called non-consumptive research and can help address this issue, such as XSample (Andresen et al. 2022) or derived text formats (Organisciak et al. 2014; Schöch, Döhl, et al. 2020).

1.3 Is representativeness possible?

For Computational Linguistics tasks, a corpus must be balanced and representative (Calvo Tello 2021). Collections of texts are created to provide access to texts and the ability to analyze them using machine learning algorithms. The creation of text corpora for analysis in linguistics and Digital Humanities has several aspects. Schöch mentions in particular the aspect of the

"population", which designates the total number of relevant items from which a corpus could be sampled or selected, given a specific scope of the corpus in terms of language, period, text type etc. (Schöch 2017a). The data selected for inclusion in a corpus can be a random selection of cases from the population, if the population is known and all texts are accessible, in which case we can speak of a *representative* corpus. Alternatively, a corpus can contain a minimum number of cases for each possible combination of values of the criteria, and may in this case be called a *balanced* corpus. Finally, a corpus can include a selection not from the population, but from a readily available source of data, in which case one may speak of an *opportunistic* corpus (Schöch 2017a; Calvo Tello 2021). Balanced and opportunistic selection are alternatives to representative sampling, because the latter is a very labor-intensive strategy, if it is realistic at all. This procedure allows for valid statements to be made about the population based on the sample and serves as a standard of comparison for other data collections. However, the population must be finite and known, which can be challenging and costly to achieve. Additionally, there must be digital or analog availability of the selected datasets and difficult decisions must be made regarding how to treat all relevant works equally or weigh them based on factors such as their distribution and reception (Schöch 2017a).

Representativeness of a corpus is being actively discussed in the scientific community. Randi Reppen points out that in some cases representativeness is possible, especially if all texts of a particular time period, author, or event can be collected, but this is rather the exception. More often than not, representativeness is governed by corpus size and material selection. Thus, smaller corpora are more often used to study grammatical phenomena and patterns, while vocabulary is more representative of larger corpora (Reppen 2022). Katherine Bode suggests that debates about corpus representativeness should focus on identifying ontological gaps and epistemological biases in evidence, and adapting editorial theories and practices from textual studies to the digital context, rather than demanding a single, "correctly balanced sample". The aim is to characterize the transformations that have produced the available evidence, rather than constructing a dataset in which all types of literature from all periods are equally represented (Bode 2020).

1.4 Research-driven or curation-driven corpus building?

In the literature we have found, we encounter different approaches to the building of a corpus. Thus, depending on whether a corpus is being assembled for the study of a linguistic or literary phenomenon or whether the goal is to assemble a collection of texts, research-driven and curation-driven approaches are distinguished.

There are many collections of texts that are actively used by various scientists as a database for their research. The list is particularly extensive for the English language. For example, corpora such as *HathiTrust*, the *Brown Corpus*, and *Project Gutenberg* are examples of widely used collections of texts for CLS. For the German language, notable examples of such a corpora are the *Deutsches Textarchiv* and the *Digital Library* in the TextGrid repository, which provide a lot of literary data. We have covered research based on these and many other corpora in the following chapters.

Sometimes creating a corpus for an individual task will be more effective than taking an already existing collection of texts. There are examples of corpora for which materials are selected according to language and genre, such as the *Corpus of Novels of the Spanish Silver Age* (CoNSSA) (Calvo Tello 2021), the *Diachronic Spanish Sonnet Corpus* (DISCO) and DISCO PAL corpus of poetic texts for Spanish (Ruiz Fabo, Martínez Cantón, and Calvo Tello 2018). Similarly, the corpus created as part of the *Riddle of Literary Quality* project covers texts published in the Netherlands, and is used in several studies of that project (Koolen 2018). Overall, there are

corpora that cover many aspects of literary research, such as time periods, many genres, one language, etc. We have also covered some of them in the further chapters on corpus building in specific scenarios.

References

See works cited and further reading for this chapter on Zotero.

Citation suggestion

Evgeniia Fileva (2023): "Introduction to Corpus Building". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/corpus-intro.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{fileva_2023_corpus,
  title = {Introduction to Corpus Building},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/corpus-intro.html},
  langid = {english}
}
```

2 Introduction to Preprocessing and Annotation

Andressa Gomide (Porto)

2.1 Introduction

A corpus is rarely a collection of raw texts grouped together. To derive meaning from corpora, we normally need to first (a) remove noise or unwanted text; (b) divide the corpus into segments (normally tokens and sentences); (c) add layers of annotation (or information) to the text; and (d) document and store information about the texts themselves (metadata). These pieces of information should be ideally structured in a way that allows them to be easily retrieved in a corpus.

The issue of annotation and preprocessing in Computational Literary Studies covers a number of aspects. The two key categories of annotation are document-level annotations (metadata) and token-level annotations (tagging). For the purposes of this survey, we deal with document-level annotation or metadata in the sections on Corpus Building, because metadata is an essential aspect of corpus design and corpus composition. Tagging can be performed at the level of individual tokens, but can also concern longer spans or regions of the text. For example, it is typical of text preparation in CLS research that text structure is annotated using markup such as XML-TEI. In addition to (and usually before) tagging texts, a certain number of preprocessing steps are typically performed, such as cleaning and normalization. Finally, preprocessing steps such as segmentation and filtering of tokens can be performed. These two latter manipulations lead over to analysis, where they are sometimes integrated and performed on-the-fly rather than being performed as part of the data preparation.

2.2 Preprocessing

Annotation is best done to a raw text, without extra information other than the text itself. However, some pre-processing to normalise the text can be desirable (Wolfe 2002; Dash 2021). In addition, if the source text already has structural annotations, for example in XML-TEI, a strategy needs to be found to preserve this information, or loosing this information be accepted. For matters of clarity, in this text, we divide this pre-processing into two groups: cleaning and data preparation.

2.2.1 Cleaning

Cleaning refers to removing noise or unwanted information from the text. These undesirable elements can be, for example, page headings and footnotes, page numbers, line breaks, sensitive data, and boilerplate text (text fragments repeated without major changes). In some cases, the cleaning has to be done manually (e.g. search for sensitive data). Very often, the cleaning

can be done (semi) automatically with the identification of mark-up and/or the use of regular expressions.

Once the cleaning is completed, the discarded data cannot be accessed anymore. This is the first version of the corpus and is normally kept as canonical raw data. It is also at this point that the information about each text (metadata) is documented; see chapter "General Issues in Corpus Building" (Chapter 1).

2.2.2 OCR and spelling

There are two additional areas where a correction of texts directly concerns problems of preprocessing and annotation: OCR errors and spelling variation. We discuss these areas as directly related to language variation in a corpus.

The consensus for dealing with OCR errors seems to be the following: if noise that alters texts is 'moderate' and uniform across a given corpus, it can be ignored. This can save a lot of resources by making perfectly proofread OCR texts unnecessary. Research that simulates uniform OCR errors shows that authorship attribution scenarios, for instance, can handle up to 20-30% of global erroneous characters without losing too much of classification strength (Eder 2013c). This finding also holds for recognized handwriting, making noisy HTR output viable for attribution studies (Franzini et al. 2018). The role of heterogeneous noise, however, is understood poorly. It can be a significant 'invisible' problem when texts are coming from large digital libraries and collections (e.g. Google books, Gutenberg, Gallica), since OCR quality may differ for different images, book formats, typesets, print quality. Even with comparable OCR frameworks, noise might be not uniform. In addition, in scenarios other than authorship attribution, noisy data is likely to be more of an issue.

Spelling normalization becomes a larger problem for Western texts the further we go back in time from the 19-20th century sources that largely had standardized grammar, spelling and typographic conventions. It is common to account for major orthographic variation (e.g. Umlauts in German and the long s in Latin-based orthographies Rebora et al. (2018b), or pre-1917 Cyrillic). In Medieval manuscript sources, the regional and individual variation in orthography dramatically increases. The problem is vividly illustrated in a study by Kestemont, Moens, and Deploige (2013) that makes a lot of effort to normalize 12th century Latin texts by lemmatizing, isolating clitics, splitting contractions and generating *possible* spellings of words to catch their outside-dictionary occurrences. At the same time, spelling normalization becomes less relevant with modern fiction, where dialectal and regional variation is seen as a conscious authorial choice and is left untouched for the analysis (Gladwin, Lavin, and Look 2015).

2.3 Data Preparation

2.3.1 Segmentation: tokens and sentences

Tokenization and segmentation are steps performed in the earlier stages of most text analyses. It is normally a fast process done using deterministic algorithms (e.g. Bird, Klein, and Loper 2009) to establish token and sentence boundaries (Grefenstette 1999).

Some examples of commonly used tokenizers are OpenNLP, spaCy and TreeTagger (Schmidt 1994). Understanding (and customizing) the segmentation process is opportune when choosing a tokenizer, as they might vary on how they account for multiword expressions and ambiguous

separators, for example. Inexact tokenization can negatively affect later processes and applications with the corpus. For instance, running a dependency parser on a badly tokenized sequence may yield errors beyond the span of the problematic token or sentence.

2.3.2 Normalization

Words might have different spellings that should be preserved or normalized, depending on the research question, especially when dealing with historical texts. When variation of orthography is not relevant, a common step is to normalize the text by establishing a standard for words with alternative spellings.

A more sophisticated way of dealing with spelling normalization is to encode both the original and the normalized or modernized version, as is possible in XML-TEI, in order to retrieve the version one is interested in as needed in the analysis step. Again, this is an area where corpus design, preprocessing and analysis interact.

Another way of normalization, which is especially done with highly inflected languages, is the annotation of the tokens with its lemma (base word form) or stem (word root), or even, in simpler scenarios, the replacement of token by the lemma or stem. Although this step is here described as a preprocessing step, a lower level of annotation is necessary.

2.3.3 Data filtering

Depending on the type of analysis that will follow, it might be necessary to remove unwanted tokens that affect the statistical analysis. This is normally done by using a stopword list to filter the texts. Stopwords, or empty words, are, usually, words commonly used in a language, such as prepositions and articles. Punctuations might also be filtered out of a corpus, especially when spoken data is concerned. However, they are crucial to identify boundaries and context, e.g. questions, exclamation, citation (Jurafsky and Martin 2023, 11). In addition, this kind of filtering can also be done based on token-level annotation (e.g. removing all but the nouns and verbs) or based on relative frequency or document frequency (e.g. removing all words present at least once in more than 90% of the documents). As mentioned above, such a filtering step is often performed 'on-the-fly' as part of the analysis pipeline.

2.4 In-text annotation

Adding extra information to the texts makes analyses easier and more precise (see chapter "General Issues in Data Analysis" (Chapter 3). Annotation can add different layers of knowledge to text. It can be related to grammar, meaning, orthography, etc. Leech (1993:275) proposes seven maxims of text annotation, from which we emphasize the following three:

- "It should always be easy to dispense with annotation and revert to the raw corpus." (1)
- "(…) the annotated corpus is made available to a research community on a caveat emptor [engl.: buyer beware] principle." (5)
- "(…) to avoid misapplication, annotation schemes should preferably be based as far as possible on 'consensual', and 'theory-neutral analyses' of the corpus data". (6)

There are many categorization schemes for the different types of annotation (see, e.g., Dash 2021). There are different layers of textual annotation and the various linguistic perspectives that can be applied to units such as tokens, sentences, paragraphs, and chapters. Gardt's schema for the analysis of textual semantics, referred by Tello (Calvo Tello 2021), includes three large groups of textual components: communicative-pragmatic frame, macrostructure, and microstructure. The communicative-pragmatic frame corresponds to the basic metadata and covers the information about the producer, the reader, the situation, and the medium. Data about genre, text type and related information are stored in the macrostructure component. The microstructure frame includes several linguistic layers such as layout, morphology, lexicology, phraseology, semantics, forms of argumentation, syntax, and punctuation (Calvo Tello 2021). The use of templates or schemas for corpus mark-up enables texts to have multiple annotations, allowing users to access the version that meets their needs (Reppen 2022). As Schöch points out, the most crucial aspect of any annotation is that an annotation scheme follows an established standard (Schöch 2017a).

For matters of clarity and to avoid theory bias, we follow The IMS Open Corpus Workbench (CWB) encoding manual (Stephanie Evert 2022; Stefan Evert and Hardie 2011)) and divide the annotation types into positional (token-level) and structural (region-level) attributes.

2.4.1 Token-level / positional

Token-level annotation is normally done automatically, attributing a value to each token in the corpus. The most common types of annotation of this kind are lemmas (the word in its uninflected form) and part-of-speech annotation (POS). Annotation of named entities (such as names of people, places, organizations or works) is also very common. Token-level annotation might also indicate richer morpho-syntactical information beyond simple POS-tags or the relationship among different tokens, as is the case when parsing a text (adding syntactic annotation).

Some common tools for token-level annotation tools are TreeTagger, spaCy, Stanza, Spark and Freeling. Among the most frequently referred standards, we can cite CLAWS, the Constituent Likelihood Automatic Word-tagging System (Garside and Rayson 1987); the EAGLE annotation guideline (Ide and Suderman 2014); the Text Encoding Initiative (TEI, Sperberg-McQueen and Burnard 1994); and Universal Dependencies, notably using CoNNL-U.

2.4.2 Region-level

Region-level annotations are particularly useful to restrict the corpus queries to specific corpus regions (see chapter "General Issues in Data Analysis", Chapter 3). Leech (1993) makes the distinction between (a) representation and (b) interpretative information that can be annotated to a text. The first refers to the structure and the form of a text. It can be sentence boundaries, pauses, words, and spellings. Interpretative, however, refers to the 'hidden' information in the text and is normally done manually by experts.

Annotations targeting textual aspects relevant to CLS and often spanning multiple tokens, like figures of speech (comparisons, metaphors, etc.) or direct and indirect speech and though representation, straddle this distinction, as they are usually annotated manually, often (but not necessarily) with the intention to provide training data to an automatic process using machine-learning. This kind of region-level annotation can also be specific to certain literary forms, like speeches and stage directions in drama, chapters and parts in novels, or verses and stanzas in poetry.

2.5 Format

The previous sections discussed the different ways of obtaining and adding information to a corpus before proceeding to the analysis. However, the way this information is stored varies.

Corpora are often prepared using XML. TEI is a frequently adopted *de facto* standard. However, there are simplified suggestions, such as Hardie (2014), which do not impose strict rules. Simplified versions of TEI might be a good choice for smaller projects, but departing from TEI does come with considerable downsides with respect to interoperability.

To make a corpus ready for query systems such as CWB and Manatee (Rychlý 2007), we often need to provide the corpus formatted in a vertical format (vrt). This means each token is represented in a line and its respective tags are added to subsequent columns (Stephanie Evert 2022). Other corpus query and analysis systems, however, are able to ingest annotated XML-TEI, for example TXM or TEITOK. The decision of how to format the corpus is normally made according to how the analysis will be performed. The chapter "General Issues in Data Analysis" (Chapter 3) deals with that.

2.6 Conclusion

As this section has hopefully shown, preprocessing and annotation of texts is a complex topic with many different aspects. Depending on the type of analysis, different levels and sophistication of annotation may be required. In addition, preprocessing and annotation interact and overlap both with corpus building and data analysis. Finally, it can be observed that increasingly, in CLS research, the combination of structural markup (such as XML-TEI) and token-level and region-level linguistic annotation is desirable and, while perfectly possible in principle, does raise challenges in particular in terms of the best sequence for the various annotation steps: most taggers do not handle existing XML markup graciously, so that the need for integration of structural XML markup and token-level annotations becomes quite challenging.

References

- See works cited and further readings on Zotero.

Citation suggestion

Andressa Gomide (2023): "Introduction to Preprocessing and Annotation". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/annotation-intro.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{gomide_2023_general,
  title = {Introduction to Preprocessing and Annotation},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Gomide, Andressa},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
```

```
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/annotation-intro.html},
  langid = {english}
}
```

3 Introduction to Data Analysis

Andressa Gomide (Porto) and *Christof Schöch (Trier)*

3.1 Introduction

Once the data has been preprocessed and annotated (see the chapter "General Issues in Preprocessing and Annotation" (Chapter 2), the linguistic and literary investigation can begin. Some common categorizations of approaches distinguish bottom-up and top-down approaches (Biber, Connor, and Upton 2007); corpus-driven and corpus-based approaches (see, e.g. Biber 2005); exploratory and focused approaches (see Partington 2009; Gries 2010; Gabrielatos 2018); and deductive and inductive approaches (Stefanowitsch 2020).

These categorizations, with some differences, distinguish two main approaches. In the exploratory approach, the research starts with a very general question, with an open-mind approach to the data [Stefanowitsch (2020); p. 61]. The focused approach is much more specifically hypothesis-drive, but it can of course also investigate hypotheses raised using the exploratory approach first. When conducting a focused investigation, the observations about the data are restricted to the elements in question.

Beyond querying corpora and establishing various characteristics based on frequency or dispersion and their comparison, statistically more advanced methods are frequently used in CLS as well. One such approach are clustering methods, another one are classification methods. Finally, it is often necessary to approach a given question, issue or textual phenomenon using human annotation, in a first step, before one can attempt to train a machine learning algorithm to identify the phenomenon in question automatically.

Data analysis is very often an incremental cycle. Refined hypotheses are derived from previous research. For this, and many other reasons, when performing data investigations, it is crucial to document and share data and procedures taken to achieve the final results.

3.2 Tools for corpus analysis

The choice of tools for data analysis varies according to the research question(s); the language used; the type of analysis; personal interest; the type of data annotation, etc. Usually, web applications like SketchEngine (Kilgarriff et al. 2014) respectively NoSketchEngine or standalone tools like AntConc (Anthony 2022), TXM (Heiden and Lavrentiev 2013) or stylo (Eder, Rybicki, and Kestemont 2016) are preferred among researchers with no programming skills, for their usability. However, such tools have some drawbacks, such as the limited number of functionalities and limited flexibility and customizability. For users looking for customization, the use of suitable programming language libraries is preferred. Examples include NLTK (Natural Language Toolkit), spaCy (a Python-based library for NLP), Gensim (a Python-based library for topic modeling and Word Embeddings), Mallet (a Java-based library for Machine Learning, including topic modeling) or scikit-learn (a general-purpose machine-learning library for Python). A downside of this method is the steep learning curve, but the effort is usually worth it. Normally,

the more sophisticated or powerful the tool is, the more it will allow the user to better explore the potential of the data.

3.3 Searching

A fundamental approach to dealing with textual data is searching in them.

3.3.1 Types of queries

Corpora can be queried by the orthographic word(s) and its variation; by the relationship of these tokens to other tokens (context); or by a combination of these two. In many cases, orthographic searches (or word form searches) are sufficient. Also, most corpus search systems allow for the use of regular expressions, which allows for searches by patterns. However, to search for advanced combinations of tokens and their annotations, we need a robust query system. These systems allow us to look for (sequences of) tokens and their annotations. Some examples are the Corpus Query Processor (CQP) (Stefan Evert 2008), the Corpus Query Language (CQL) (Jakubíček et al. 2010), and ANNIS (Zeldes et al. 2009).

3.3.2 Precision and recall

As powerful as a query processor may be or as sophisticated as a query may be written, it is very unlikely the search result is a perfect representation of what we intended. We have to consider that among the results we can have true positives (what is relevant) and false positives (not relevant). We should also consider that search might fail to identify some relevant results (false negative) and identify some irrelevant results (true negative). With data on these four cases, we can calculate the precision (ratio of relevant retrieved items and all retrieved items) and the recall (ratio of retrieved relevant items over all relevant items) of the results.

3.4 Frequency

Another fundamental aspect when dealing with textual corpora in CLS is to consider the frequency of word tokens or other features.

3.4.1 Significance testing and effect-size

Frequency is the core of quantitative studies, but frequency on its own can be deceiving and unreliable, as it only gives information about the sample our data is aiming to represent. For this reason, statistical significance tests are frequently used to verify whether a result was not obtained by chance, but that it "reflects the characteristics of the population from which the sample was drawn" (Sirkin 2006, 306). Effect-size measures can also be used to obtain the strength of the difference or relationship found in the results (see, e.g. Ellis 2010, 3–5; Gabrielatos 2018).

3.4.2 Contrasting

A common way of investigating the difference between two datasets or two subgroups of texts within a corpus is through the calculation of keyness, or distinctiveness. By calculating keyness we can identify keywords. The usual understanding of keyness is that a word or other feature is key when it occurs more often in a corpus or corpus section than would be expected based on a reference or comparison corpus. However, semantic definitions of keyness referring to aboutness, salience or discriminatory power have also been proposed (Schröter et al. 2021) and may motivate the use of dispersion in addition or instead of frequency to define keyness measures. For a better understanding of different measures to calculate keyness or distinctiveness and their implications, see Du et al. (2021).

3.4.3 Dispersion

Besides the aforementioned tests, another important piece of information is to know how the observed pattern is distributed across corpus units. This is normally acquired by applying dispersion measures. Dispersion may be defined as the degree to which occurrences of a word (or lemmas, phrases, annotations of any sort) are uniformly distributed across a corpus. If a word occurs much more often in one text of a corpus than in the other texts, it can be said to be unevenly dispersed. Conversely, an evenly dispersed word is expected to have a relatively constant presence across all corpus texts (Gries 2008). There are different methods and measures to calculate dispersion (see e.g. Gries 2021). A very simplistic, yet widely used way of reporting the distribution of frequencies in a corpus is the document frequency, as it takes only a simple count of how many texts or sections in the corpus feature the searched word.

3.4.4 Cooccurrences and relationships

Investigating the relationship between elements (tokens, annotation) in texts is a very common practice when analysing corpora. A straightforward method is the generation of frequency lists of n-grams. N-grams are sequences of a given number of words. They are widely used in language modelling (see, e.g. Jurafsky and Martin 2023). In research focused on the discourse, these sequences are also known as lexical bundles (Biber 1999).

A commonly used technique to investigate cooccurrences for specific words is the exploration of collocations, or "the phenomenon surrounding the fact that certain words are more likely to occur in combination with other words in certain contexts" (Baker 2006). The method used to identify collocations impact greatly on the results and the choice for the most suitable measure varies according to the intent of the research (see, e.g. Stefan Evert 2009).

3.5 Supervised versus unsupervised methods

Supervised learning (also called classification) is the machine learning term for problems in which we use computational aid to give us answers about new data, teaching the computer not just to recognize specific patterns, but also to relate such patterns to specific, previously-defined categories, such as authors. In order to do so, applying supervised methods requires dividing the data into training and test datasets, with the first serving as model data which the algorithm uses to learn the patterns. This means of course, that the training dataset needs some kind of labels (e.g. author names if we are training for authorial signal detection, or in more advanced cases marking specific patterns that are to be learned) that will indicate specific distinctive

groups. Contrary to the supervised methods, unsupervised learning (also called clustering) does not require division into two datasets – the algorithms of this kind immediately group texts based on their specific features (e.g. words, or chunks of words or characters, or more complex semantic, syntactic, or structural features).

3.5.1 Clustering

Within CLS, there is a range of occasions where clustering methods are commonly used. The most prominent example is probably authorship attribution, where dendrograms (tree diagrams) based on text similarity matrices, in turn based on text similarity measures, are routinely used to display and discuss results. In addition, Principal Components Analysis is frequently used for stylometric analyses, whether focused on authorship or on other categories such as gender, literary period, or literary genre or subgenre. One of the interesting features of PCA is that the weights of features associated with each principal component can also be inspected, something that often supports interpretation. A prominent tool supporting the creation of distance-based dendrograms and frequency-based PCA plots is stylo (Eder, Rybicki, and Kestemont 2016).

When we want to have a general understanding of the main ideas in a corpus, clustering and topic modeling are frequently used techniques of data exploration. Clustering involves grouping the tokens in the corpus into meaningful groups. Two common clustering methods are the scatter browsing method and the hierarchical clustering. The scatter method is more suitable for an exploratory approach (see e.g., Cutting et al. 2017; Hearst and Pedersen 1996), while the hierarchical allows for query-specific investigations [Feldman and Sanger (2006); pp. 82-84].

Topic modeling is another approach that can be understood as being based on identifying clusters of words based on their co-occurrence patterns. It is often used to aid researchers to summarize and explore large corpora: "Topic modeling algorithms (…) analyze the words of the original texts to discover the themes that run through them [and] how those themes are connected to each other" (Blei 2012, p 7). Latent Dirichlet Allocation (LDA) (also described by Blei, Ng, and Jordan 2003) is currently the most used model, but many alternatives exist.

Very briefly put, topic modeling can be understood to be a process of dimensionality reduction. Using the in-document co-occurrence of words as the key information, the often very large term-document-matrix is decomposed, in a sense, into two matrices of much smaller size: a term-topic matrix and a topic-document matrix. Because there are usually much fewer topics than there are word forms, this yields not only a dimensionality reduction, but also an analytically-useful summary of the data. Two prominent tools supporting topic modeling are Mallet, a Java-based command-line-tool, and Gensim, a Python library.

An interesting, recent alternative to LDA-based topic modeling is the use of large language models to discover clusters of semantically-related words. One such approach is available in the Python-based tool BERTopics that uses a combination of one of many transformer-based language models and TF-IDF calculations to derive word clusters.

3.5.2 Classification

Finally, classification is another fundamental approach in CLS research. It is ubiquitous in CLS and NLP. Many of the automatic methods for token-based or region-based annotations described in the chapter on "General issues in preprocessing and annotation" (Chapter 2) are in fact classification-based sequence labeling approaches.

Classification as a key branch of machine learning is also used in many other applications in CLS research. Again, authorship attribution (see Chapter 5) is the most prominent example,

where classifiers are trained to recognize the most likely author of an unseen text based on the properties of a large number of texts of known authorship. However, this is just one example, and classification is used for any number of tasks relevant to literary studies: whether to identify the class of a given token (as in Named Entity Recognition or metaphor detection), to classify specific sequences of texts (for example in the context of direct / indirect speech and thought recognition), or to classify entire texts by any number of categories (such as author gender, subgenre, time period, canonicity status, and many more).

3.6 Conclusion

It is not possible to give appropriate space to each and every method of analysis relevant to CLS research in this introductory text. However, within the different perspectives of this survey, the different chapters devoted to several key domains of interest within CLS attempt to provide a closer survey of current practices of analysis in the field.

References

See works cited and further readings for this chapter on Zotero.

Citation suggestion

Andressa Gomide, Christof Schöch (2023): "Introduction to Data Analysis". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/analysis-intro.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{gomide_2023_analysis,
  title = {Introduction to Data Analysis},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Gomide, Andressa and Schöch, Christof},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/analysis-intro.html},
  langid = {english}
}
```

4 Introduction to Evaluation

Christof Schöch (Trier)

4.1 Introduction

Evaluation, in the context of CLS research, means the process of and procedures for verifying the quality of the method of analysis in terms of its performance, accuracy, robustness or generalisability. Within the workflow model of research we follow here, evaluation is a step that becomes especially relevant towards the end of the workflow, after the analysis has been run. However, multiple cycles of annotation, analysis and evaluation may of course be required before any given study is completed.

Note that evaluation as described here, primarily concerned with assessing the *performance* or *accuracy* of an analytical process, is different from evaluating the *validity* of a method (or the quality of the operationalization), where the key question is how well a chosen measure or method is really indicative of the phenomenon that one intends to study. This consideration about the "fit between a measurement that is applied and the theoretical construct it aims to represent" (Herrmann, Jacobs, and Piper 2021, 472) is sometimes discussed under the heading of construct validity.

Evaluation is a key element in the research process in CLS. One may argue that one of the core features of Computational Literary Studies is that it proceeds by using formalized, algorithmic and/or statistical methods to investigate literary texts and literary history represented in the form of digital data. Some of this research and the methods it uses are exploratory, in which case evaluation is in large part a matter of plausibility and contextualization of results, but some exploratory methods (such as clustering) also support formal evaluation. Other research and the methods it uses are based on the paradigm of hypotheses-testing and classification, in which case formal evaluation is probably the dominant mode. In addition, but usually prior to the application of exploratory and/or hypothesis-driven methods of analysis, automatic or manual annotation of the data may be performed, and needs to be evaluated as well.

Given the current state of the art in evaluation, very much focused on measuring the agreement of annotators or measuring the performace of machine learning algorithms, we can structure the issue of evaluation in CLS research into several main areas briefly described in the following section and taken up, as appropriate, in the different chapters of this survey that are concerned with evaluation.

4.2 Research and evaluation scenarios

4.2.1 Evaluation of manual annotations

Manual, qualitative annotations can be evaluated, in the absence (usually) of a gold standard, by calculating the inter-annotator agreement (sometimes also called inter-rater reliability) using

a range of measures intended for this purpose. Measures used for this purpose in CLS include Cohen's Kappa, Fleiss' Kappa and Krippendorff's Alpha (see Fleiss 1971; Artstein and Poesio 2008) or the more recently proposed Gamma (Mathet, Widlöcher, and Métivier 2015).

4.2.2 Evaluation of automatic annotation

The evaluation of automatic, token-based annotations that have been applied in a classification-based approach, can usually be evaluated against a gold standard or reference annotation. Such a reference annotation is usually required for training of the algorithms. Examples that have not been used in the training phase can then be used for evaluating the resulting performance. This performance is measured in terms of true and false positives and true and false negatives, based on which indicators such as precision and recall, and further derived scores such as the F-score, can be calculated. A very useful introduction to the evaluation of classification-based methods is the chapter 8.5.1, "Metrics for Evaluating Classifier Performance", in (Han and Kamber 2012).

4.2.3 Evaluation of classificatory approaches beyond annotation

Evaluation in the context of classification-based methods other than automatic token-level annotations, for example document-based classification tasks, are not fundamentally different to automatic token-level annotations. Again, the performance can be measured in terms of true and false positives and true and false negatives, based on which indicators such as precision and recall, and further derived scores such as the F-score, can be calculated. An error analysis based on a confusion matrix can in many cases yield important insights into the structure of the classification problem. Finally, it can be useful to determine which features have had the strongest influence on the trained model's decision-making, for example by extracting feature weights from a model.

4.2.4 Evaluation of clustering methods

In evaluation of clustering methods when a gold standard is present, such measures as adjusted Rand Index or cluster purity are usually used. A useful introductory chapter on this topic is chapter 10.6, "Evaluation of Clustering", in (Han and Kamber 2012).

4.2.5 Evaluation of exploratory approaches other than clustering

Evaluation of other exploratory methods, such as topic modeling or keyword extraction, can be done either using qualitative and subjective methods of evaluation (such as establishing the plausibility of results by comparing them to results from earlier research), or using downstream-tasks that can, for example, be framed as a classification problem, in which case the metrics mentioned above for this scenario can be used. Examples for scenarios using such a downstream classification tasks include Schöch (2017b) and Du, Dudar, and Schöch (2022).

4.3 Further forms of quality assurance

There are of course other aspects of evaluation, or quality assurance more generally, that apply to research in CLS. Apart from traditional forms of peer review, it is worth mentioning at least two aspects: the issue of replication / reproducibility, closely linked to Open Science principles, and the issue of code, data and tool review and criticism.

4.3.1 Replication and reproducibility: an emerging issue

Research strategies that include enabling and/or performing replication, reproduction or other forms of follow-up research should be mentioned here, because they increasingly become an issue not just in Artificial Intelligence and Natural Language Processing, but also in CLS research Such strategies can indeed be used to assess the quality, coherence, robustness and/or reliability of earlier research, in particular when performed by others. The demands on documentation as well as on data and code availability are rather high, however, if they are to truly support replication and/or reproduction. As a consequence, this aspect of CLS is still in an emerging stage (see Huber and Çöltekin 2020; Schöch 2023).

4.3.2 Data, code and tool review

There is one further aspect of evaluation and quality assurance in CLS research, in some way connected to the previously mentioned issues of replication and reproduction. What is meant are critical approaches and reviewing practices applied not just to traditional scholarly publications, but also to other forms of scholarly output that are crucial to research in CLS. For example, digital editions and corpora or other datasets are regularly being reviewed, for a number of years already, in RIDE – A review journal for digital editions and resources. Similarly, datasets can be described, whether in short form or with examples for applications in research, at the peer-reviewed Journal for Open Humanities Data (JODH). Journals such as the *Journal for Computational Literary Studies* are taking first steps towards implementing code and data review into their peer review process.

Finally, tool criticism is another aspect of these critical practices. Traub and van Ossenbruggen define tool criticism as "the evaluation of the suitability of a given digital tool for a specific task" (Traub and van Ossenbruggen 2015, 1). Traub and van Ossenbruggen stress that the primary objective of tool criticism is to "better understand the bias of the tool on the specific task", and that improving the tool itself is a secondary concern (Traub and van Ossenbruggen 2015, 1). When taking tool criticism to actual research practice, however, its implications become increasingly concrete.

4.4 Conclusion

Evaluation of methods is an important aspect of research in CLS and, depending on the approach used, takes various different forms. There is definitely a trend to more formal and more reflected evaluation in CLS across the various approaches and research issues that goes beyond the area of classification, as becomes visible from the entries on evaluation in specific areas of research in CLS.

References

See works cited and further reading for this chapter on Zotero.

Citation suggestion

Christof Schöch (2023): "Introduction to Evaluation". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/evaluation-intro.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{schoch_2023_evaluation,
  title = {Introduction to Evaluation},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Schöch, Christof},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/evaluation-intro.html},
  langid = {english}
}
```

Part II

Authorship attribution

This part of the survey is devoted to issues of stylometric authorship attribution.

5 What is Authorship Attribution?

Joanna Byszuk (Kraków)

5.1 Introduction

Authorship attribution can be simply explained as a study whose goal it is to answer the question: who is the author of some text we examine?

Nowadays, non-traditional, computational or stylometric authorship attribution or stylometry for short – as it is also often called to distinguish it from more long-standing, philological methods of authorship attribution – is one of the more popular tasks in Natural Language Processing and Computational Literary Studies or even a subdiscipline on its own, but its roots are deep in the past. The questions "is this text authentic? was it really written by this person?" and "I wonder who could have authored this unsigned or otherwise anonymous text" have probably been accompanying humans since the beginning of written communication. In fact, some of the early approaches to authorship attribution that we know of use methods very much like the ones we apply today. For example, in 1851, Augustus De Morgan suggested that Biblical authors could be identified by the features of their writing, which was also used in early stylometric ventures dedicated to the study of the chronology of texts, verifying whether a text was falsified, etc.

A very similar method was used by Frederick Mosteller and David Wallace in their study (Mosteller and Wallace 1963) of *The Federalist Papers*, a famous collection of essay pamphlets published by a trio of American forefathers: Alexander Hamilton, James Madison, and John Jay, to promote the new Constitution. Using frequencies of selected words from the essays and analyzing them with a Bayesian-based method, they identified the author of twelve Papers, the authorship of which had been disputed earlier. While their method had its faults and better ones have been developed since, Mosteller and Wallace are and will be remembered as pioneers in computational authorship attribution.

5.2 Authorship Attribution or Verification?

There are two distinguishable types of authorship examination – attribution and verification (Koppel, Schler, and Argamon 2009).

Authorship attribution typically concerns the so-called closed set problems, that is all inquiries in which the real author must be one of the finite, known, set of candidates. Think of *The Federalist Papers* as the perfect example – a collection of 85 articles and essays written by three founding fathers whose names are well known to us: Alexander Hamilton, James Madison, and John Jay. While an extremely skeptical person could argue it is impossible to know if they didn't have another anonymous co-writer unknown to the history, no evidence points to such a scenario and it is safe to assume that 12 letters of long disputed authorship must have been written by one of these three. In the case of closed sets, the focus is on distinguishing between

individual candidates' stylistic fingerprints and finding out which of them best fits the disputed text.

On the contrary, authorship verification usually concerns the so-called open set problems, that is all inquiries in which there is some suspicion over possible authors, usually based on philological evidence, but the possibility that the real author is not actually included in the corpus cannot be excluded. This kind of situation can result from many factors, such as many possible writers (especially in under-researched problems), little data available, and finally, the possibility that no other texts by the real author are preserved (or that the author completely hid themself under a pseudonym) and so cannot be included in the reference corpus.

While both have been applied in numerous stylometric investigations, attribution obtains significantly more attention in applications (starting with Mosteller and Wallace 1963) and review of methods (Grieve 2007; Stamatatos 2009; Stefan Evert et al. 2017) than verification, with the latter approach developing more intensely only in the last decade or so (e.g. Koppel and Schler 2004; Kestemont et al. 2016b; Halvani, Winter, and Graner 2019).

5.3 Tools for Authorship Attribution

The most widely-used tool in stylometric authorship attribution is stylo (Eder, Rybicki, and Kestemont 2016), as it provides users with the choice between a simple graphical user-interface and a command-line interface, implements most currently-used methods of stylometric authorship attribution (many of which can also be used in applications beyond authorship attribution), is continually developed and has been popularized through more than a decade's worth of introductory and advanced workshops by the developers. Another well-established tool for authorship attribution is JGAAP (Java Graphical Authorship Attribution Program).

References

See works cited and further readings on Zotero.

Citation suggestion

Joanna Byszuk (2023): "What is Authorship Attribution?". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/what-author.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{byszuk_2023_authorship,
  title = {What is Authorship Attribution?},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Byszuk, Joanna},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
```

```
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/what-author.html},
  langid = {english}
}
```

6 Corpus Building for Authorship Attribution

Artjoms Šeļa (Kraków)

6.1 Introduction: Corpus building and language variation

Corpus building is crucial step in many authorship attribution studies, because it is primarily through corpus design that researchers exercise control over a multitude of factors that might influence textual differences. This control is necessary to attempt to isolate the authorial signal on the grounds of linguistic evidence. Research on language variation (e.g. Tagliamonte 2011; see also Grieve 2023) says that this variation is simultaneously driven by many things, and tends to be nested: we can detect differences between large groups of speakers (dialects), but also between different regions of the same group, different cities within a region, different neighborhoods within a city. Similar effects were observed multiple times empirically in written texts: given a large corpus, we can detect overarching genre splits (e.g. poetry vs. prose), period and gender differences, register differences, and, finally, authorial differences that can be additionally organized by social background and education of authors. It lead some researchers (C. Labbé and Labbé 2001; D. Labbé 2007) to seek one 'similarity scale' for text analysis that would allow to tell apart possible influences at different values of distance scores. It is, however, impossible, to define such a scale universally for the multitude of relevant methods, languages and corpora. Instead, the dominant strategy in practice, especially in method evaluation research, is control of the corpus. Below we outline three areas where corpus preparation is most visible and discussed: (1) chronology, (2) genre & register and (3) document size and corpus sampling. In the end, we discuss alternative approaches to handling language variation in a corpus.

6.2 Chronology

Language is continuously changing and each generation of writers adapts a different version of it. Texts written in similar periods will tend to naturally group together, and stylistic distances between texts will tend to increase with the distance between them in time (Juola 2003; Hughes et al. 2012), with strong generational effects visible as well (Underwood et al. 2022). This creates a natural problem of difference inflation in chronologically unbalanced corpora. As Patrick Juola put it (Juola 2015, 106): "A collection of blog posts in modern English would not provide an adequate control sample for Elizabethan. […] Marlowe would still probably have written Shakespeare's plays if the alternatives were […] bloggers for the New York Times'. In practice, it is common to limit the background corpus by the narrow time period of a text / corpus in question (e.g. Grieve 2007; Forsyth and Holmes 2018; Rebora et al. 2018a; Kocher and Savoy 2018).

Additionally, styles of individual authors tend to have their own chronological dynamics: they change over time and this change is traceable by stylometry. In fact, early European stylometrists at the end of 19th century gathered around the problem of dating Plato's dialogues rather than for attribution goals (Grzybek 2014). Studies also demonstrate that change of writing styles in

authors are unequal: stylistic drift might be very pronounced in some cases (David L. Hoover 2007), but it is far from being universal (Reeve 2018; Nagy 2023). The existence of stylistic drift makes a case for distinguishing early/late works in corpus preparation for authorship attribution, or calls for nuanced approaches to sampling authorial style across time (Barber 2018) in order to reduce the regular effect of chronology.

6.3 Genre and register

It was many times observed that literary genre — at different levels of definition — has systematic effects on stylistic differences: fiction is distinct from non-fiction (Piper 2017), poetry differs from prose (Chaudhuri et al. 2018), tragedy differs from comedy (Schöch and Riddell 2014). These effects can be united under the linguistic notion of 'register': a specific situation under which a text was produced and addressed to a specific audience (Biber 1988). The practice of corpus normalization for 'register' is uneven and differs in its resolution. The 'narrative fiction' is often considered to be a uniform register (at least silently) in attribution and evaluation practice (e.g. Eder 2013c; Stefan Evert et al. 2017; Kocher and Savoy 2018), but there is an obvious linguistic variation within fiction, too, for example as an effect of narrative perspective. Many studies strictly limit registers, e.g. working only on letters addressed to one person (Tuccinardi 2016), only on Latin historiographies (Kestemont et al. 2016a) or only on newspaper columns (Grieve 2007). This illustrates the nested nature of linguistic variation for which a fully 'controlled' corpus potentially becomes an endless process of zooming into specific conditions and constrains under which a group of texts was produced.

A special case of the difference in register is poetry in verse that linguistically is very distinct from prose. The major driver of this effect is non-arbitrary division of speech into periods (lines) and poetic meter that, by strictly governing prosody, systematically (re)organizes natural language. As a result, poetic texts in verse radically differ from prose even on the basis of a handful of most frequent linguistic features (Chaudhuri et al. 2018; Storey and Mimno 2020). This poses a significant obstacle for any study that deals with mixed corpora, or even with early modern drama that regularly mixes prose and poetry (and proportions of this mix can be different across dramatic genres).

In addition, different meters twist language in different ways; the form is a major confound of any quantitative claim made about poetic texts. Thus, major evaluation studies (Plecháč 2021; Nagy 2021) always work on metrically homogeneous corpora.

6.4 Size and sampling

One of the crucial questions of corpus design for authorship attribution is the question of size. How large the texts should be to be 'large enough'? Usually this is asked in relation to an established 'nearest neighbors' methodology that is based on measuring distances between frequency distributions of most frequent features (words, character n-grams). Empirical tests offer a range of answers: from 5000 words as a minimum in a scenario where all quantified texts are of the same size (Eder 2013b), to 1000-2000 words in a scenario where a much larger reference or training corpus is available (Eder 2017). In practice, researchers often work with texts as small as 500-2000 words. Effective sample size depends on the nature of used features and can be decreased further when instead of linguistic 'rare events' it uses features that are densely distributed even in short texts: these can be derived from the formal structure of text (poetic rhythm, phonic organization of rhyme, see Plecháč 2021; Nagy 2021). Alternatively, a range of derived summary statistics (like vocabulary richness, complexity etc.) can also be used

effectively in combination with frequency-based vectors (Weerasinghe, Singh, and Greenstadt 2021).

The question of size, again, cannot be answered universally; even the stylistic recognizability of individual authors is not something even or constant (Eder 2017) and some might require larger samples than others. Forensic authorship attribution often deals with extremely short texts (letters, notes) and many frequency-based methods are simply not applicable in this domain. The whole analysis paradigm differs as research resorts to various word n-gram tracing methods (Nini 2018; Grieve et al. 2019), or non-linguistic attribution (graphology, material evidence).

The question of document size is part of a larger problem of uneven text representation. In the vast majority of cases, stylometry is not taking samples of style from authors in controlled conditions, but is sampling the historical record of style that is, more often than not, uneven. The multivariate comparison of large texts alongside small texts introduce biases and can provide unreliable results (Moisl 2011). The common solution to this problem is using various sampling strategies, like 'downsampling' a corpus to a smallest text to make any comparison even (oversampling, that instead generates synthetic data points to match larger samples, is exceedingly rare (e.g. Feng and Hirst 2013)).

Moisl (2011), however, approached the problem of document length from a direction of behavior of individual features. He introduced a screening method that determines minimal 'reliable' sample size for each of the used features (the original study was based on English bigrams). Based on a normality assumption and the properties of a binomial distribution, he adapted a sample size function that estimates (in a given confidence threshold) a size of a document *required* to reliably represent a probability of a given feature. This allows to non-arbitrarily cull the corpus and feature space, balancing the number of features vs. the number of documents used. In turn, this can escape employing bootstrapping or extensive sampling strategies and can be useful in unsupervised scenarios (Cafiero and Camps 2019), but is currently under-tested.

It is also common to combine smaller texts with known authorship into even samples (Plecháč 2021; Rebora et al. 2018a, 2018b). Random samples that combine different sources can be particularly effective, since they draw stylistic information that is not stratified by any artificial condition (e.g. of one text, one chapter, one theme). Eder (2013c) has shown for 19th century English novels that random sampling that evenly draws from the available bag of words performs much better than consecutive samples, or 'chunks' of novels, which may be too dependent on local context to represent style effectively.

6.5 Beyond the corpus control

In a authorship attribution evaluation study, Jack Grieve writes that "the anonymous text is the product of a single situation and so each author-based corpus should be composed of texts produced in the most similar register, for the most similar audience, and around the same point in time as the anonymous text" (Grieve 2007, 255). To obtain such a fully controlled corpus is increasingly demanding: one can add the preference for texts being of similar length, or being produced, typeset, machine recognized under similar conditions. John Burrows noted that uniform corpus representation in many cases is unrealistic and will always come with information trade-off: "we need to determine what is most appropriate, accepting only such limitations as we must, and resisting them when we can" (Burrows 2007, 46). There is also concern of generalizability of methods: if most evaluation studies are made using highly specific slices of language, then how we can be sure they will perform similarly in any other scenario? At the extreme level of granularity each text can be viewed as its own 'register', written under

conditions that never existed before and will not exist again. Here, a 'laboratory' control of textual conditions becomes an unreachable utopia.

The concerns and pressures of appropriate corpus construction created a sub-branch in stylometry of cross-genre or cross-register authorship attribution with corpora tinkered specifically to be heterogeneous (Kestemont, Mike 2012; Barlas and Stamatatos 2020; Wang, Xie, and Riddell 2021). Here linguistic variation is taken as a methodological challenge, not solely as an issue of corpus design. For example, 'unmasking' techniques (Koppel and Schler 2004) test the stability of authorial signal by iteratively removing most distinct features, that might drive surface differences between samples (i.e. induced by different genres or themes). Attempts have been made to also 'control' for various stylistic influences *a posteriori*, penalizing genre or chronological signals in a corpus (Calvo Tello 2017; Underwood and So 2021). Authorship verification problems that ask about the likelihood of sample X coming from author A are conceptually defined as 'one-class' classification problems, even if in practice, they are usually set up as multi-class (Halvani, Winter, and Graner 2019). The minimal context required to solve a verification question is just writings of one author and there are notable attempts to remove the control corpus (and 'distractor authors') from analysis altogether, and just focus all inference on the information coming from stylistic behavior of same-author texts (Noecker Jr and Ryan 2012; Halvani, Winter, and Graner 2019). In the end, we see that the language variation which fuels the concern around the corpus in stylometry, can be also either embraced and recognized, or bracketed out by radically minimizing the surrounding context.

References

See works cited and further readings on Zotero.

Citation suggestion

Artjoms Šeļa (2023): "Corpus Building for Authorship Attribution". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/corpus-author.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{sela_2023_corpus,
  title = {Corpus Building for Authorship Attribution?},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Šeļa, Artjoms},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/corpus-author.html},
  langid = {english}
}
```

7 Annotation for Authorship Attribution

Julia Dudar (Trier)

Generally speaking, stylometric authorship attribution relies primarily on un-annotated texts, simply using word forms as the fundamental feature. However, even this requires tokenization to be performed. In addition, some studies, especially when working with short texts, aim to use more information than just word forms for stylometric analysis.

Please note that the remarks on OCR and spelling normalization that are explained in the chapter "General issues in Preprocessing and Annotation" (Chapter 2) are also relevant for authorship attribution. Indeed, if texts written by different authors come from different sources, they may have undergone systematically different editorial preparation, including spelling normalization or modernisation, something that can interfere with authorship attribution.

7.1 Tokenization

In most studies doing authorship attribution, only relatively basic techniques of processing the string of characters are used, notably tokenization in order to identify individual word forms. The reason for this is that most stylometric authorship attribution operates at the level of individual word frequency analysis. It is an issue for ideogrammatic scripts such as Chinese, where one word (in the sense of a semantic unit that would be translated into one word, usually, in alphabetic languages) consists of (typically) 1-3 characters and boundaries between words are not marked by spaces.

7.2 N-Gram creation

In only a relatively small number of studies, n-grams at the word or character level have been used as an alternative to single word forms (unigrams). One prominent supporter of n-gram methods for authorship attribution is Brian Vickers. He has authored a number of works dedicated to authorship attribution for Early Modern Drama using the rare 3-6-word n-gram matches between a corpus of plays of a target author and a play of unknown author as a method (Vickers 2008, 2011, 2012). His main hypothesis is that word-n-grams are more suitable for authorship attribution tasks compared to single tokens because authors tend to choose semi-constructed word combinations (Vickers 2011).

The use of word n-grams for authorship attribution was also proven to be useful by Antonia, Craig, and Elliott (2014). For their analysis, the authors evaluated two n-grams methods: "strict n-grams", which is based on the use of straightforward sequences of 1-5 word n-grams and "skip n-grams", which involves the omission of function words. As corpora they used a collection of English Renaissance plays and a collection of articles written for Victorian periodicals. The findings of their analysis revealed that 3-grams obtained the highest results in various evaluation scenarios for both corpora when using the strict n-gram method. In contrast, when using the skip n-gram method, the 4-grams demonstrated the best performance.

However, other authors do not support this idea and have obtained different results using the n-gram method (see Craig and Kinney 2009; Jackson 2010; David L. Hoover 2015). David Hoover (2015), for instance, replicated tests on the Vickers method using 3- to 6-word-n-grams on Victorian drama. His results demonstrated that frequent single words were the most effective indicators of an author's style, while the Vickers method proved unsuccessful according to his experiment. Subsequently, he expanded his research (David L. Hoover 2018) to assess not only rare 3- to 6-word n-grams, but also character-2grams, -3grams, and -4grams, as well as word-2grams, -3grams, and -4grams. The outcomes revealed that frequent single words still yielded the best results. Although character-2grams performed relatively well at times, they still produced inferior results compared to frequent words. Longer n-grams sequences displayed much weaker results for both character and word n-grams. Similar study but in a cross-language setup was conducted by Eder (2011).

López-Escobedo et al. (2013) combined functional and content word n-grams with various other features (see following sections). They extracted both types of n-grams: straightforward sequences and sequences with a gap. Concerning the length, they used unigrams, bigrams and trigrams.

7.3 Lemmatization or POS-Tagging

Even less frequently, linguistic annotation is used. There are some instances where lemmatization is employed (C. Labbé and Labbé 2001; Eder and Górski 2023). Sometimes, POS-tagging is used and usually combined with n-gram calculation, because POS usually reduce the number of features to a very small number and using n-grams multiplies the number of different features. Also, POS-tagging can be used to disambiguate word forms, as is routinely done by Hugh Craig in the "Intelligent Archive" (Craig and Whipp 2010). Besides word n-grams, López-Escobedo et al. (2013) also performed authorship attribution combining POS-n-grams with other features. The length of POS-n-grams varied between 1-3-n-grams.

7.4 Syntactic annotation

In some rare cases, more sophisticated linguistic annotation, including dependencies or other syntactical information, is used in stylometric authorship attribution. Cinkova and Rybicki (2020) developed an approach that enables the comparison of an original text and its translation to identify an author's style across languages. They constructed a parallel corpus of Czech-German literary texts and enriched the texts with Universal Dependencies, a framework for consistent annotation of parts of speech, morphological characteristics, and syntactic dependencies across diverse languages. Despite this, the method's effectiveness remained limited. After that, the authors subsequently added a shared pseudolemma with each annotated lemma in both languages, and it resulted in a clear performance improvement, pushing performance up to 95.6%.

7.5 Other features

Equally rarely, additional features beyond standard linguistic annotation are used in authorship attribution. An example is Jacobs and Kinder (2020), who also used stylistic features such as type-token ratio as well as content-oriented features such as sentiments. Another rare case is Suzuki et al. (2012), who have used co-occurence-based features.

An interesting study was conducted by Gómez-Adorno et al. (2018), where the authors evaluated a variety of stylometric features categorized into three groups: phraseological, punctuational, and lexical. It is important to note that they avoided using typical stylometric features such as word frequency and instead incorporated features like lexical diversity, average word length, average sentence length, standard deviation of sentence length, average paragraph length, standard deviation of paragraph length, document length, and punctuation marks. Such additional stylometric features including the distribution of word length frequencies, type token ratio or Hapax legomena count were also adressed by López-Escobedo et al. (2013).

7.6 Conclusion

Summarizing the provided information, we can conclude that although function or content word frequency is a popular, reliable and well-performing feature for authorship attribution, it may in some contexts be beneficial to incorporate additional features into the research design. Depending on the language of the corpus and authorial style, word n-grams, POS n-grams, or even syntactical dependencies can aid in identifying unique authorial styles or other previously unseen characteristics in literary works.

References

See works cited and further readings on Zotero.

Citation suggestion

Julia Dudar (2023): "Annotation for Authorship Attribution". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/annotation-author.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{dudar_2023_annotation,
  title = {Annotation for Authorship Attribution},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Dudar, Julia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/annotation-author.html},
  langid = {english}
}
```

8 Analysis in Authorship Attribution

Joanna Byszuk (Kraków)

As mentioned in the chapter "What is Authorship Attribution?" (Chapter 5), computational studies into authorship are divided into attribution and verification problems. Both of these areas make use of machine learning methods (especially supervised and unsupervised learning, that is, classification and clustering methods) that have long proved their usefulness across numerous scholarly fields.

8.1 Attribution versus verification

To summarize the distinction between attribution and verification explicated above, "authorship attribution includes determining which of the candidate authors in the examined dataset is most likely to be the author of the text in question. In turn, authorship verification deals with checking whether any of the candidate authors is at all likely to be the author of the examined text" (Hernández-Lorenzo and Byszuk 2022).

A good preparation for authorship attribution must be preceded by a thorough study of relevant literature – as many of the texts of disputed authorship have been subject to literary and philological (even palaeographic) investigations that serve as an invaluable source of information on possible authors. It is also even more crucial than in other cases that open set authorship studies include the step of authorship verification next to (or before) the step of authorship attribution.

8.2 Authorship attribution methods

Apart from selecting the supervised (classification) or unsupervised (clustering method) – see chapter "General Issues in Data Analysis" (Chapter 3) for details – and the type and number of features, many methods of authorship attribution analysis requires considerations about the choice of distance measure (i.e. a similarity measure projected onto a multidimensional geometric space) to examine stylistic relation between the texts, an area that has seen a lot of research devoted to it (see e.g. Grieve 2007; Argamon 2008; Eder 2015; Stefan Evert et al. 2017). Relatively little work has been done on the performance of various features, with the evidence pointing to the most frequent words as the most reliable carrier of the authorial style signal (Eder 2011; Camps, Clérice, and Pinche 2020). For more information on the issue of the features, see the chapter "Annotation for Authorship Attribution" (Chapter 7).

Authorship attribution studies frequently use various unsupervised machine learning methods for exploratory data analysis. Most common are clustering approaches, including bootstrapped cluster analyses visualized in the form of networks, in stylometry also bootstrap approaches, but also used are methods of dimensionality reduction, such as Principal Component Analysis (PCA) and Multidimensional Scaling (MDS). The above-mentioned are included in the most commonly used R package 'stylo', but as these are standard statistic methods, they and others can also be found in more general packages in libraries in both R and Python.

There are relatively few examinations comparing the performance of particular methods. Most work in the field focuses on cluster and network approaches, and this area was given some focus in Ochab et al. (2019) who tested the performance of typical methods of cluster and network grouping, finding Ward linkage method and Louvain method of community detection the most reliable.

While in many simpler cases clustering will be a good enough method for authorship attribution, more precise classification methods give better certainty in cases where the authorial signal is weak or blurred, or in the case of shorter texts. Supervised machine learning classification with SVM (support vector machines), Delta method and NSC (nearest shrunken centroids) is often considered more reliable. As hinted above, in this type of analysis, one of the above-mentioned classifiers 'learns' the style of each of the authors based on which knowledge it is able to point which of them the text of disputed authorship is the most similar to.

Benchmark comparisons of machine learning methods for authorship attribution are also relatively outdated, with the most notable, being Stamatatos 2009 (listing dozens of studies describing methods) and a study of Jockers and Witten (Jockers and Witten 2010) proposing NSC and RDA (regularized discriminant analysis) as best performing. Jockers and Witten criticize SVM at length, however, this method is often used and recommended by computer science experts such as Stamatatos, Koppel, and Argamon, and has proven to be more reliable and stable when dealing with high dimensional and sparse data, such as historical and shorter writings (Stamatatos 2013; Franzini et al. 2018), and two- and three-class problems (Luyckx and Daelemans).

Savoy (2020b, 109, reformulated here) distinguishes the following steps of a quantitative inquiry, that can also be applied to attribution:

- Defining a precise research question or formulating a hypothesis.
- Preparing a selection of texts that will make up the evaluation corpus. (See also: section "Corpus building for Authorship Attribution")
- Preprocessing data to ensure its quality (which can include normalization of spelling and/or removing extra-textual elements such as page numbers).
- Choosing a text representation strategy, that is which stylistic features are to serve as a marker of style.
- (optional) Removing noisy attribute features to lower computational costs or improve method performance.
- Choosing and applying a machine learning algorithm to perform the classification.

In the case of authorship attribution problems, the question of the right selection of texts that are to make the corpus is particularly important.

Once we have formulated our hypothesis and identified all texts that might relevantly represent possible (candidate) authors, we proceed with dividing them into a training set and a test set. The texts included in the training set will serve as the learning data for our classifier, so should best reflect the style of particular authors. The test set will include the investigated text as well as other texts by all authors represented in the training set. The inclusion of candidate authors both in the training and the test sets is aimed at helping us measure the performance and reliability of the classification. If our classifier has learned to recognize particular candidate authors, say with 80-90% accuracy (although different values will be considered 'good' for a various number of classes in an experiment), we know that it has a fairly good idea of how particular authors write and what stylistic features distinguish them – this allows us to trust that it recognizes the author of the investigated text similarly well.

8.3 Authorship verification methods

Authorship verification deals with the question of whether any of the candidate authors is at all likely to have written particular text. The main difference from the attribution approaches described above is that in this case, rather than try to guess the author, the algorithm compares pairs of texts against the others to see whether any of them is significantly more similar to one another than to the rest of the dataset.

The best-described approach is the General Imposters (GI) framework, first proposed by Moshe Koppel (2004; Koppel and Winter 2014) and further examined and developed by Mike Kestemont (Kestemont et al. 2016b), Patrick Juola (Juola 2015), and others. As explained in Kestemont et al. (2016b), "[the] general intuition behind the GI, is not to assess whether two documents are simply similar in writing style, given a static feature vocabulary, but rather, it aims to assess whether two documents are significantly more similar to one another than other documents, across a variety of stochastically impaired feature spaces (Eder 2012; Houvardas and Stamatatos 2006) and compared to random selections of so-called distractor authors (Juola 2015), also called 'imposters' " (Kestemont et al. 2016b, 88).[1]

To put this procedure in simple terms, it is based on running a series of classifications comparing each text in the dataset to the examined text (Target) and to a random and changing subset of candidate authors (Imposters). The final outcome is a value between 0 and 1, showing how much each text was closer to the Imposters (0) and Target text (1). Importantly, apart from the general 'the higher the better attribution accuracy confidence' value, the framework includes calculating "Periods of Confidence", that is, dividing the 0-1 range into three parts: definitely not Target text, cannot say for sure, definitely Target text.

8.4 Applications of authorship attribution

As far as applications of stylometric authorship attribution methods are concerned, they are of course legion, given that the field has been active at least since the late 1950s, with important precursors far into the 19th century.[2] Early and pionieering work using modern statistics from the late 1950s includes Grayston and Herdan (1959), Ellegård (1962), Mosteller and Wallace (1963) or Levison, Morton, and Wake (1966). Today, such applications can be found for cases of disputed or uncertain authorship in a wide range of languages, periods, and literary genres; from the wide range of applications, we can only list a somewhat arbitrary selection (ordered simply by year of publication):

- C. Labbé and Labbé (2001) and Cafiero and Cafiero and Camps (2019) analysed French drama of the seventeenth century (the Molière-Corneille case);
- Binongo (2003) uses Principal Component Analysis to find out who wrote the 15th book in the Wizard of Oz series (written primarily by Frank L. Baum);
- van Dalen-Oskam and van Zundert (2007) investigate questions of authorship and scribal influence of the Middle Dutch Walewein;
- Craig and Kinney (2009) investigated several authorship cases concerning English drama of the late 16th and early 17th century (around Shakespeare, Marlowe and others);
- Jannidis and Lauer (2014) investigate German 19th-century novels;
- Juola (2015) describes the process by which Joan K. Rowling's authorship of *A Cuckoo's Calling* was established;

[1] Cited after https://computationalstylistics.github.io/blog/imposters/.

[2] For an extensive list of early and later applications, see the Stylometry Bibliography curated by Christof Schöch since 2016 and containing around 3500 entries.

- Rißler-Pipka (2016) conducts experiments regarding authorship attribution of Spanish novels of the Early Modern period;
- Kestemont et al. (2017) investigate several hypotheses regarding the authorship of the *Wilhelmus*, the Dutch folk song that is the national anthem of the Netherlands;
- Tuzzi and Cortelazzo (2018) and Savoy (2020a) built suitable corpora and use several authorship attribution methods to attempt to identify the true author behind the contemporary Italian bestselling novels published under the name of Elena Ferrante;
- Grieve et al. (2019) dealt with a mid-19th-century American-English letter (often attributed to Lincoln, but attributed to Hay by Grieve and colleagues);
- Mazurko and Walkowiak (2020), developed stylometric methods for examining authorship of literary texts in Ukrainian;
- Hadjadj and Sayoud (2021) used over-sampling and PCA to deal with authorship attribution challenges in imbalanced corpora in Arabic;
- Ai et al. (2021) applied an LDA-Transformer model to perform Authorship Attribution of Chinese Poetry;
- Vega García-Luengos (2021) worked on a corpus of plays by Spanish writer Lope de Vega and proposed new attribution hypotheses;
- Most recently, Jungmannová and Plecháč (2022) investigated 20th-century novels in Czech (regarding Milan Kundera).

This list, though far from complete, hopefully conveys at least some of the richness of methods and domains of authentic authorship problems addressed with computational approaches.

8.5 Conclusion

In conclusion, it is worth noting that, while many methodological questions remain open in stylometric authorship attribution, it remains one of the oldest and most developed areas of investigation within Computational Literary Studies.

References

See works cited and further readings for this chapter on Zotero.

Citation suggestion

Joanna Byszuk (2023): "Analysis in Authorship Attribution?". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/analysis-author.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{byszuk_2023_analysisauthorship,
  title = {Analysis in Authorship Attribution},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Byszuk, Joanna},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
```

```
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/analysis-author.html},
  langid = {english}
}
```

9 Evaluation in Authorship Attribution

Joanna Byszuk (Kraków)

Conducting a reliable and reproducible study in authorship attribution can be challenging, and the topic has gathered a lot of attention over the last years. For a long time, authorship attribution studies relied on rather simple methods of evaluation – reporting classification accuracy for known cases in the examined corpora, or using two or three different methods to ensure the same result is reached with various means. Most of the papers cite the specific settings (number and type of features, classification algorithm and distance measure) that were used. Many studies also mention determining the best settings for the particular corpus, before using those settings to classify a text of unknown authorship.

The evaluation of authorship attribution is done not only in application papers that attempt to solve actual attribution problems but also in methodological papers that attempt to validate certain approaches or understand particular parameters. Some such examples include studies mentioned in the previous section, examining how particular methods perform in attribution tasks (e.g. Jockers and Witten 2010; Stamatatos 2009) or what are the limitations of the methods when it comes to the length of the samples or language (e.g. Rybicki and Eder 2011; Eder 2013b).

One of the biggest problems for reproducibility is the lack of access to corpora or proprietary code. While nowadays it is a standard to openly publish both, e.g. on Github or Zenodo, unless they are under strict copyright, for many older studies it is impossible to find the exact corpora used, or even the full description of their content that would facilitate recreating.

The panel at the Digital Humanities Conference 2020 by Schöch, Van Dalen-Oskam, et al. (2020) provided the most comprehensive discussion yet of the issues related to the reproducibility of computational literary studies. Contributions ranged from developing a typology of repeating research (distinguishing features of replication, reproduction, etc.), difficulties with funding and conducting replications of old studies, and scarcity of evaluation methods.

9.1 Methods of Evaluation

9.1.1 Evaluation in the case of clustering-based authorship attribution

Clustering approaches do not have a straightforward way of evaluating the results, therefore studies employing such methods usually rely on good practice rules for conducting reliable study. As noted in Eder (2013a), "[a] vast majority of methods used in stylometry establish a classification of samples and strive to find the nearest neighbors among them. Unfortunately, these techniques of classification are not resistant to a common mis-classification error: any two nearest samples are claimed to be similar, no matter how distant they are." This is particularly characteristic of clustering and network approaches, as they are based on grouping the elements of the dataset based on similarity (finding the said nearest neighbors), and will always find some connection between all elements, even if they actually have little in common. Therefore, corpus design is of crucial importance – failing to include probable authors will produce a

misattribution, and including too many candidates (of whom some could not have authored the text for objective reasons such as being already dead or not yet born) can introduce noise that will make the conditions of the classification more difficult or even impossible.

As Maciej Eder further argues, clustering approaches are very sensitive to the number and type of features used, which can and usually do influence the results. While in the case of strong and highly distinctive authorial signals the differences might be minor, they can nevertheless lead to erroneous conclusions. To counter this risk, Eder proposes applying bootstrapping (a technique used across various scholarly fields employing statistics), a procedure in which a series of experiments is performed, with each varying in the number of features used, e.g. ten experiments increasing from 100 to 1000 most frequent words in 100-word intervals. The results of each round are then put together and only texts that were recognized as each other's nearest neighbors in as many cases as the set threshold (usually 50%) are considered reliably close. The method has since gained popularity in authorship attribution circles, and is now commonly applied in most studies.

Stefan Evert et al. (2017) compares the performance of various distance measures across feature vectors 100, 1000 and 5000 words long, using (1) the difference between z-transformed means, (2) the Adjusted Rand Index, and (3) clustering errors, finding that Cosine Delta produced the most accurate results and longer vectors usually resulted in fewer errors. It should be noted however, that for each corpus a different number of features might be enough to produce a reliable result, as observed in Eder (2017) and efforts are made to adjust the methods and the selection of frequencies to work on shorter texts Eder (2022).

9.1.2 Evaluation in the case of classification-based authorship attribution

Classification approaches to authorship attribution are evaluated in all studies. While relying on good practices described in methodological and benchmark papers is of course present, most, if not all, methods, provide classification accuracy scores with the results – these, however, take various forms, from accuracy, precision, recall, F1-score, or AUC (Area Under Curve), all traditionally used in machine learning.

A straightforward and often applied method of evaluation is cross-validation, and in particular the *leave-one-out* method, in which a series of experiments are performed, each time taking one text out of the corpus and comparing it against the remaining ones. This allows to verify the success rate of the classification both in the overall and in the detailed scopes, and to identify texts that are misclassified or introduce noise to the corpus.

More complex evaluation procedures include testing the performance of various classification methods and types of features before coming to conclusions (e.g. Grieve 2007).

9.2 Conclusion

In conclusion, one can say that authorship attribution is probably the domain within CLS that has the most developed practice of evaluation. The reason for this is not just that it is a very long-standing domain, but also that there are sufficient numbers of undisputed, single-author publications to make systematic evaluation, and evaluation-driven development of methods, feasible. In addition, authorship is – despite edge-cases of various modes of collaboration – a category with much clearer class boundaries than in the case of canonicity (which can be understood as a gradual attribute) or genre (where one text can participate in more than one genre, to various degrees). As a consequence, evaluation in authorship attribution is, arguably, more feasible than in the case of categories such as genre or canonicity.

References

See works cited and further readings for this chapter on Zotero.

Citation suggestion

Joanna Byszuk (2023): "Evaluation in Authorship Attribution?". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/evaluation-author.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{byszuk_2023_evaluationauthorship,
  title = {Evaluation in Authorship Attribution},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Byszuk, Joanna},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/evaluation-author.html},
  langid = {english}
}
```

Part III

Genre Analysis

This part is devoted to issues of the analysis of literary genre.

10 What is Genre Analysis?

Christof Schöch (Trier)

10.1 Introduction: The theory of literary genres

Literary Genre Analysis investigates literary texts from the perspective of the genre(s) or subgenre(s) they participate in. As a consequence, genre analysis in CLS operates against the backdrop of both an extensive body of theories of genres and subgenres in established literary history, and a no less extensive body of studies on the characteristics and the history of specific literary genres and subgenres.

Arguably, the nature of and distinctions between literary genres is one of the oldest problems of literary theory, going back at least to Aristotle, who thought of genres as categorical classes that are mutually exclusive and clearly delimited based on specific criteria. For example, narrative and dramatic forms are distinguished based on whether or not the speaker is a part of the represented world or not. Based on such classificatory thinking, hierarchical systems of genres and subgenres were pursued, where each class has a *differencia specifica* relative to other, related classes.

Many different perspectives on genre have been developed by literary theorists and literary historians over time. As Matthew Wilkens writes, mentioning several key perspectives: "For critics in the mold of Northrop Frye, genre is a formal category having to do with the 'radical of presentation' through which plot-level events are portrayed; for Gérard Genette, it involves primarily content-level differences within formally defined modes; for Raymond Williams, genre is part of a "social language" that unites distinct aspects of the social and material processes that make up a cultural situation" (Wilkens 2017, 2).

Starting in the 1970s, in addition, the categorical approaches were questioned more radically and alternative and more flexible models of literary genre were proposed (see e.g. Hempfer 1973, 2014). Prominent examples include, for instance, genres as groups of works linked by *family resemblance* (in the tradition of Wittgenstein), or genres as *prototypical* categories (in analogy to prototype theory in biology and cognitive psychology). These approaches have in common that an individual work can participate in a given subgenre to a larger or lesser degree, and can participate in more than one genre to varying degrees. Both José Calvo Tello and Ulrike Henny-Krahmer have pursued the consequences of this kind of conceptualizations of genre in their research (see Calvo Tello 2021; Henny-Krahmer 2023).

Another major issue in the theory of genres that has significant relevance for computational approaches to genre is the question of whether genres manifest themselves as inherently textual properties of texts, or whether they are rather socially-constructed phenomena. This has direct implications for the question of whether or not approaches focusing only on textual properties rather than on contextual information are suitable for investigations of literary genre.[1]

[1]This is an issue that has clear parallels to the situation in research on canonicity and prestige, see the chapter "What is Canonicity?" (Chapter 25).

As far as the description, theory and history of specific genres is concerned, this is quite obviously a large, important and dynamic domain within literary studies. Studies on specific genres and subgenres – such as the novel, the short story, the science fiction novel; such as drama, comedy or tragedy; the sonnet, etc. – are plentiful, whether in a specific literary period, across several periods, or in a theoretical perspective. Also, because individual genres can be defined and described at various levels – among them theme, style, form, setting, protagonist, plot or audience – investigations of all of these aspects can be relevant in a genre analysis. To give just some examples: genre is often defined at the level of formal, structural or stylistic characteristics, such as poetry in verse compared to narrative prose or dramatic dialogue interspersed with stage directions. However, theme is also a major aspect of genre and subgenre, as in the distinction between the romance novel (primarily about love and obstacles to love) as opposed to the science fiction novel (primarily about imaginary alternatives, based on scientific and social innovations, to human existence as we know it). Personnel, of course, can also play a role, as in the kinds of characters that are typical of tragedies as opposed to comedies. Then, setting can be relevant, as in the urban settings typical for crime fiction as opposed to the outer limits of civilization characteristic of many adventure novels. Many more examples for relevant aspects could be added here.

For research in CLS, studies covering such aspects in various genres not only provide essential background information, but also raise research questions and offer hypotheses to investigate based on metadata or corpora.

10.2 Computational Genre Analysis

Against this background, there are two levels at which different kinds of literary genre analysis can be distinguished, based on a distinction between various degrees of granularity: (a) At the level of the texts to be analyzed, where entire documents, chapters, paragraphs, pages or even sentences can be the object of attention. And (b) at the level of the genre categories, where broad genres (such as the distinction between narrative and dramatic texts), more specific subgenres (such as the distinction between comedy and tragedy, adventure novel and science-fiction novel), or even more fine-grained distinctions within subgenres (such as the distinction between detective fiction and hard-boiled crime fiction) may be investigated. Most research in CLS is focused, nowadays, at the second level of granularity.

Another key distinction is whether the investigation of genre (a) starts from known genre assignments of texts and aims to discover features (of various types) that are characteristic of texts of each genre category (contrastive analysis), or whether (b) establishing the association between texts and genre categories is the principal task (e.g. classification tasks), or whether (c) any structure in the data that can be discovered is then checked for correlations with genre categories (e.g. clustering tasks). Mixed approaches are frequent, as there are investigations using some form of regression analysis.

Finally, a large proportion of research is interested in genre not so much in order to attribute texts to genres or subgenres, or to identify general characteristics of given genres or subgenres, but is rather focused on texts belonging to one particular genre or subgenre category, where this category is accepted *a priori*. The focus in this case is rather to analyze some specific literary phenomenon within a corpus defined by its genre or subgenre. Additional research is either based on the results from such classification tasks, or uses alternative approaches to such classification tasks.

Further details on the corpora used for genre analysis, on the kinds of annotations used, and on the methods of analysis used in selected research belonging to these different categories are provided in the following sections.

References

See works cited and further readings for this chapter on Zotero.

Citation suggestion

Christof Schöch (2023): "What is Genre Analysis?". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/what-genre.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{schoch_2023_genre,
  title = {What is Genre Analysis?},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Schöch, Christof},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/what-genre.html},
  langid = {english}
}
```

11 Corpus Building for Genre Analysis

Evgeniia Fileva (Trier)

11.1 Introduction

The key challenge in corpus building for genre analysis is the composition of the corpus in terms of the genres or subgenres targeted, on the one hand, and in terms of competing or potentially interfering categories such as authorship, period or formal characteristics, on the other. Often corpora used for research in Computational Linguistics or Digital Humanities (such as stylometry) are designed from the start to include only one genre or subgenre (e.g. newspaper articles, encyclopedias, theater plays, tragedies, etc.), although the actual level of homogeneity of such corpora can be a matter of debate. For such tasks, the primary genre-based classification of texts is important because texts belonging to the same genre (or class) are assumed to demonstrate similar linguistic features.

Often, however, classifying texts by genre, or in a more fine-grained manner by subgenre within a broader genre, particularly for the purpose of creating a corpus, presents a separate, specific problem. The main reasons for this are that genre categories are hard to define in a coherent, systematic way; that sensible categorizations of genres are very much bound to the respective literary tradition (language, period) they apply to; and that a single text may participate in multiple genres at the same time (see the chapter on "What is Genre Analysis?" (Chapter 10) for more information). As a consequence, Calvo Tello (2021), for instance, approached the problem of genre classification as a "multi-level classification task". Overall, for genre-based research it is important that corpora contain specific information about genre stored in metadata.

11.2 Corpora frequently used for genre analysis

11.2.1 Curation-driven, general-purpose corpora

Typically but not necessarily, curation-driven, general-purpose corpora are relatively large in volume and relatively heterogeneous with respect to the genres or subgenres represented in them. However, there are also curation-driven corpora for CLS that focus on one particular, larger genre, such as drama, narrative or poetry. In addition, it is typical for curation-driven corpora to be monolingual and for texts contained in them to be encoded in XML-TEI, although exceptions to both points certainly exist.

The *Brown Corpus* is among the oldest, well-established corpora used for genre-specific tasks in particular in Computational Linguistics. Conceptually and practically, the Brown Corpus is constructed very differently from, for example, the vast Hathi Trust Library, and consists of 500 samples each of about 2,000 words, with a total of 802 texts, all first published in 1961, representing a range of styles and varieties of prose (see Brown Corpus Manual and Francis and Kucera 1979). The corpus is not meant to represent standard English, but rather a standardized body of data for comparative studies. For his study, Kessler, Nunberg, and Schutze (1997) used only

499 out of 802 texts from the Brown Corpus, which were selected using a custom classification system. The texts were analyzed based on three categorical facets: Brow (levels of intellectual background required by the target audience), Narrative (binary, whether a text is written in narrative mode), and Genre (reportage, editorial, scitech, legal, nonfiction, fiction). The corpus was divided into a training subcorpus (402 texts) and an evaluation subcorpus (97 texts) for the study, which were selected to have roughly equal numbers of all represented combinations of facet levels. The texts in the evaluation subcorpus were chosen using a pseudo random-number generator, resulting in different quantitative compositions of the training and evaluation sets, with some genre levels being more frequent in one set than the other (Kessler, Nunberg, and Schutze 1997). Then an analysis of each text in the evaluation set was conducted.

Closer to current concerns in CLS research, such corpora as *Deutsches Textarchiv* (DTA), *Corpus of Novels of the Spanish Silver Age* (CoNSSA) or *Théâtre classique* are also monolingual corpora. Thus, the DTA covers the period from 1600 to 1900 and includes about 4,400 works in total, of which 1,500 belong to a balanced core corpus and about 700 are literary works ('Belletristik'). As stated in the corpus description, the collection was created to reflect the diversity of the German language, so the collection is characterized by genre richness as well as scrupulous supervision in the process of digitizing the material. Furthermore, the corpus website offers a researcher-oriented navigation tool, the "linguistic search" within the DDC (Dialing/DWDS-Concordancer) linguistic search engine. For each text, DDC creates a machine-readable index file containing additional information for each word that can be used during queries. These index files are created from inputs in a DDC-specific XML format, which contains all the information available to DDC but is not readily efficiently queryable. The developers suggest using the DTA as a reference corpus for linguistic research.

The *Project Gutenberg*'s text collection counts texts in more than 60 languages, all open access and public domain. The project description indicates that in addition to books there are also such units as manuals, pamphlets, periodicals, travelogues, theses, journals, or chapbooks. One of the features of this corpus is that instead of genres it uses "categories", that is, topics by which books are sorted. The official website states the following description: "The collection includes eBooks on many topics. There is emphasis on literary works and reference items of historical significance, because volunteers have focused on digitizing such works. Any eligible item, on any topic, is welcome." (Gutenberg).

The most widely used text collections for German are the *Digitale Bibliothek* in the TextGrid Repository and the *Deutsches Textarchiv* (DTA). The TextGrid Digital Library offers a comprehensive collection of XML/TEI-encoded texts from fiction and non-fiction literature from the beginning of book printing until the first decades of the 20th century, written or translated into German. The collection is of particular interest to German and comparative literary studies as it contains almost all important canonical texts and numerous other literature-historical relevant texts whose copyright protection period has expired. The texts are mostly from reliable editions and are therefore citeable. The TextGrid Repository makes these texts available to the general public not only for reading but also for further processing, such as in editions and corpora. The XML files were converted into a valid TEI format, which allows for precise searching and analysis. The metadata only contains some very broad genre specific information such as verse, prose and drama. Both corpora have, for example, been used in the research of Trilcke, Fischer and colleagues (Trilcke, Fischer, and Göbel 2016; Trilcke, Fischer, and Kampkaspar 2015) among other corpora such as Wikisource and Projekt Gutenberg-DE. The authors describe the DTA corpus as having high-quality TEI markup, but containing relatively few texts. The German-language branch of Wikisource also has limited texts. The *Projekt Gutenberg-DE* archive has poor markup with only basic XHTML. The TextGrid Repository, which contains basic TEI markup, is the most applicable option in their view (Trilcke, Fischer, and Kampkaspar 2015).

The *European Literary Text Collection* (ELTeC) belongs to the group of curation-driven corpora that have a clear focus on one particular literary genre, in this case the novel. ELTeC consists of a number of sets of novels in different European languages, where each set contains 100 novels first published in the period 1840 to 1920 in one given language-based literary tradition. ELTeC is not a representative corpus, but uses a number of corpus composition criteria to ensure that the variety of production in each literary tradition is represented in each corpus. These criteria concern aspects such as author gender, text length, degree of canonicity and publication time (Schöch et al. 2021). However, ELTeC provides virtually no metadata on the subgenres of the novel represented in each collection, because of the challenges connected to establishing a taxonomy across multiple literary traditions. ELTeC is a multilingual (more than 17 European languages) collection consisting of 12 complete corpora of individually selected 100 novels, with the corpora being comparable to each other in their internal structure, size, and composition. ELTeC also contains additional corpora of various sizes. All metadata (title, author, publication date, etc.) are modeled according to the same standard and encoded in XML-TEI, using a set of ELTeC-specific schemas (Burnard, Odebrecht, and Schöch 2021).

Théâtre classique is a collection of French dramatic texts edited by Paul Fièvre since 2007, with (currently) 1,700 French plays, the majority of which were written or published between 1630 and 1800 (see Fièvre 2007). The platform provides contextual information and several statistical and analytical perspectives on the textual data. The XML-TEI data contains good document-level metadata and detailed structural markup typical of dramatic works (acts, scenes, stage directions, character speeches), while the the TXT and HTML files lack much of this level of detail. The corpus has been described as "an essential enabling force for recent, quantitative approaches to French drama" (Schöch 2018a) and is now available also via the DraCor platform.

11.2.2 Research-driven corpora built for specific purposes

Sometimes new corpora are created for specific tasks. Such an example of the creation of the original text collection can be seen in Calvo Tello (2021). *The Corpus of Novels of the Spanish Silver Age (CoNSSA)* was created on the basis of Spanish-language literary prose that was written between 1880 and 1939 be authors from Spain. The corpus includes novels by 107 authors who met all the criteria that have been set for that corpus by Calvo Tello: novels in Spanish by Spanish authors published between 1880 and 1939. In addition to basic bibliographic information, Calvo Tello collected the following metadata about authors and their works: name (full and short), years of life, author's preferred genre, number of pages dedicated to the author in the *Manual de literatura española* (MdLE), and provenance information. Calvo Tello states that categorizing genres and subgenres have been a particularly challenging task. The author proposes a method of encoding information of unprecedented detail about genre and subgenre that takes into account, for instance, the hierarchical levels and sources of information, including super-genre, genre, subtitle, subgenre from different sources such as literary histories, editorial information, and annotations by Calvo Tello himself. The aim is to capture the complexity of genre and subgenre categorization and to reflect the blending of different subgenres in many texts. Another corpus designed specifically for the investigation of subgenres of the Spanish-language novel, but this time for novels from Argentina, Cuba and Mexico, is the *Corpus de novelas hispanoamericanas del siglo XIX* (conha19) edited by Ulrike Henny-Krahmer who also curated a bibliography as a basis for the corpus building process (see Henny-Krahmer 2017) and describes the process of modeling, building and encoding the corpus in Henny-Krahmer (2023).

As Ruiz Fabo, Martínez Cantón, and Calvo Tello (2018) argue, there is a lack of research on poetic corpora, especially for Spanish. A good example of a poetry corpora widely used in the research community is *Diachronic Spanish Sonnet Corpus. DISCO* consists of 2677 sonnets in

Spanish from the 19th century written by 685 authors from Spain and Latin America. The corpus is intended to provide a wide sample, inspired by distant reading approaches, and being updated with additional sonnets from other centuries. The poetic texts were extracted from Biblioteca Virtual Miguel de Cervantes and encoded in XML-TEI P5 format. The metadata, stored in the TEI-Header, include year of birth and death, place of birth, and gender. The corpus is available on GitHub, saved in Zenodo, and includes VIAF identifiers to enhance the corpus's findability in the linked open data (Ruiz Fabo, Martínez Cantón, and Calvo Tello 2018).

DISCO PAL was created on the basis of the DISCO corpus and is available for data mining tasks on Spanish poetry, and particularly for obtaining the Global Affective Measure (GAM) of poetry (Barbado et al. 2022). While *DISCO* doesn't provide information directly usable for text modeling tasks, (Barbado et al. 2022) presented DISCO PAL, the Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels, which is a subset of the DISCO corpus. DISCO PAL includes 274 sonnets in Spanish from different time periods annotated with affective, lexico-semantic, and psychological labels and aims to make poetry available as machine-readable data for linking, indexing, and extracting new information. While DISCO includes the metadata only about authors, sonnet scansion, rhyme-scheme and enjambment, the DISCO PAL corpus includes binary labels for psychological concepts and integer values for affective and lexico-semantic features and also provides a rich source of data for text mining tasks on Spanish poetry (Barbado et al. 2022).

Another example of a poetic corpus is the one presented by Šeļa, Plecháč, and Lassche (2022). Their research employs five poetry collections in various languages such as Czech, Dutch, English, German, and Russian to compare and analyze metrical types used in poetry.

The *EmoTales* corpus (Francisco et al. 2012) is a specific corpus designed for narrative applications. The corpus focuses on fairy tales due to their explicit representation of emotions and suitability for the identification and study of emotions. It includes 18 tales of different lengths, written in English, with a total of 1,389 sentences and 16,816 words, chosen to cover a broad spectrum of styles by having tales from different authors and time periods. It contains in-context sentences with emotional tags based on subjective human evaluations.

Another example of a research-driven corpus is the corpus developed by the Zeta and Company project corpus. This corpus consists of French novels (both written in French and translations into French) from the second half of the 20th century, currently with 1200 novels in total. The corpus has several special properties. First of all, it is specifically designed for the investigation of several popular subgenres of the French novel, namely science fiction, crime fiction and sentimental novels, in comparison with a wide range of highbrow novels. For this reason, the corpus is balanced with respect to these four groups, within each decade (as far as feasible) as well as across the period 1950-2000. Second, the subgenre information is derived from the bibliographic metadata about publishers and collections in which the novels were published, under the assumptions that specific collections (such as Harlequin's 'Duo Romance' or Gallimard's 'Série noire') publish only novels that can be categorized as sentimental novel or crime fiction novel, respectively. And third, the novels all being still under copyright, the digitisation and usage in the research project is covered by the 'Text and Data Mining-Exception' that applies to research purposes in German copyright law. As the full texts cannot be published, the project makes available a subset of (currently) 320 texts as a so-called 'derived text format' that allow statistical calculations, but make reading the novels impossible (see Organisciak et al. 2014; Schöch, Döhl, et al. 2020).

11.3 Conclusion

Thus, building a corpora for genre analysis in Computational Linguistics and Digital Humanities could be a quite challenging task. The composition of the corpus in terms of targeted genres and classifying texts by genre or subgenre present specific challenges since genre categories are hard to define, and a text may belong to multiple genres. Curation-driven general-purpose corpora tend to be large and heterogeneous, whereas research-driven corpora focus on the selection of research material.

References

See works cited and further readings for this chapter on Zotero.

Citation suggestion

Evegniia Fileva (2023): "Corpus Building for Genre Analysis". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/corpus-genre.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{fileva_2023_corpusgenre,
  title = {Corpus Building for Genre Analysis},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/corpus-genre.html},
  langid = {english}
}
```

12 Annotation for Genre Analysis

Christof Schöch (Trier)

12.1 Introduction

Apart from document-level, bibliographic metadata, which we describe in the chapter on "Corpus Building for Genre Analysis" (Chapter 11), the issue of annotation for genre analysis includes, in the perspective of this survey, two major areas: (a) the manual identification or automatic extraction, from literary texts available as digital full-text, of document-level information relevant to genre; and (b) the annotation, again manually or automatically, with information relevant to genre, of tokens or spans within the literary texts.

What kinds of information, however, are relevant to genres or subgenres? As we have seen in the chapter "What is Genre Analysis?" (Chapter 10), the answer to this question can be derived from the information that is relevant to the definition and/or description of literary genres and subgenres in the literary practice, as established by literary studies. As it turns out, a great variety of aspects are relevant to the genre or subgenre of a text, among them form, style, theme, personnel, setting, mode of publication, or audience. The question of determining which features are best suited for genre analysis, or most relevant for specific genres, is a matter of debate within CLS. In addition, however, computational approaches can (and frequently do) simply use the frequencies of word forms as a feature for approaching genre, in which case only a minimal annotation in the sense of tokenization is required.

Finally, as noted also in the chapter on "Corpus Building for Genre" (Chapter 11), we of course find studies that annotate and analyse either corpora of texts containing one specific genre to be characterised more closely through analysis, or corpora containing several distinct genres or subgenres intended for a contrastive analysis or in order to develop methods that are able to distinguish between texts of different genres and classify them accordingly.

Annotation for genre analysis, then, in Computational Literary Studies, consists primarily in collecting relevant information (other than bibliographic metadata) about literary texts, whether at the document level or within texts, whether regarding form, style, theme, personnel or setting (and others), and making this information available for the analysis step described in "Data Analysis for Genre" (Chapter 13). In some cases, this step could also be described as feature engineering or feature generation, rather than annotation or tagging in the conventional sense.

12.2 Minimal annotation for genre analysis

In many cases, the analysis of genres or subgenres operates with simple wordforms and a bag-of-words model. This appears to be particularly the case in studies that use methods such as regression, clustering or classification, and when large collections of texts are analyzed. In this case, the identification of word forms (or lemmas) characteristic of a given genre or subgenre is either a goal in and of itself, or the features derived in this way are the input for a subsequent cluster, regression or classification-based analysis.

A classic of this kind of analysis is Kessler, Nunberg, and Schutze (1997), albeit in a perspective more of Computational Linguistics than CLS. This paper formulated the influential idea of distinguishing between *generic facets* (abstract properties of texts relevant to genre) and *generic cues* (surface features indicative of specific generic facets). The authors also defend the idea of simple, surface features to be just as useful for (broad) genre classification tasks over more complex features representing structure or content.

Other examples of a study using virtually no annotation (other than bibliographic, document-level metadata) include Underwood et al. (2013), Schöch and Riddell (2014) or Worsham and Kalita (2018). Underwood et al. (2013) analyze a very large collection of texts, namely 469,000 volumes from the Hathi Trust Library, with the aim of generating page-level genre assignments. The authors used a statistical method, the Wilcoxon rank-sum test, to identify subsets of words from the overall vocabulary for use in various genre-oriented classification tasks. Worsham and Kalita (2018) analyze a subset of the Gutenberg corpus that they call the *Gutenberg Dataset for Genre Identification* that contains 3577 texts that can be assigned to one of six types of fiction. They use a classification approach based on deep learning and, programmatically, perform no preprocessing other than tokenization and a reduction of the vocabulary to the 5000 words that are most frequent in the corpus overall.

Schöch (2018b) used a measure of distinctiveness first proposed by John Burrows, Zeta, as a way to derive words that are characteristic of several subgenres of French drama, namely comedy, tragedy and tragi-comedy. Other than tokenisation and lemmatisation, no preprocessing or linguistic annotation was used. The resulting lists of characteristic words for each subgenre, however, were used as a basis for a subsequent cluster analysis.

Similarily, Du, Dudar, and Schöch (2022) use a variety of distinctiveness (or keyness) measures in order to obtain lists of words that can be understood to be typical or characteristic of a certain number of subgenres of the French contemporary novel. In their pipeline, which also includes tokenization, lemmatisation and POS-tagging, the information extracted is not actually collected at the document level, but only at the level of the novels aggregated into groups defined by their subgenre.

12.3 Genre information at the document level

There is an increasing number of studies, in recent years, that generate document-level annotations (in addition to bibliographic metadata) for genre analysis.

Because of the importance of the semantic level for genre analysis and genre distinctions, topic modeling is in fact one of the methods of choice for genre analysis. An early example for this is the influential monograph of Jockers (2013), in which he used topic modeling to investigate trends and distinctions in a large corpus of English-language novels from North America and Ireland.

Another example is Wilkens (2017), who analysed 8500 20th century American novels in English for their subgenre assignments using topics in combination with other features. Taking up the distinction between properties of genres and textual cues that allow to measure the prevalence of these properties in each novel, he first extracted the following features: "1. Subject matter, measured in the present case by topic-modeled word frequencies. 2. Style, form, and diction, measured by volume-level statistics including reading-level score, verb fraction, text length, etc. 3. Setting and location, assessed via geolocation extraction and geosimilarity measures. 4. A limited range of extra-textual features, including publication date and author gender" (2017, 6).

In a similar manner, Hettinger et al. (2015) have generated features for each novel in a corpus of 1700 German novels. The features concern common stylometric features (such as word frequencies), but also other features concerning topics (using topic modeling) and character interaction (using character network data).[1]

Also operating at the level of entire documents, but not using topics as features, is the study by Coll Ardanuy and Sporleder (2014). The authors have created social networks of characters in novels and then derived feature vectors from these networks to characterise novels at the level of character-based structural properties. They then use these for cluster analysis with a perspective on authorship and genre. Falk (2015) is another study generating character network features for genre analysis, but this time for a set of dramatic works. A series of studies on dramatic character networks has been published by the DLINA (Digital Literary Network Analysis) group, with Trilcke et al. (2016) in particular using network data to investigate, among other things, different kinds of German dramas, for example "open" and "closed" dramas.

Finally, a recent study by Nicholas D. Paige (2020) employs an unusual approach, in which the analysis of digital full texts plays virtually no role and instead, a considerable number of properties of a large selection of French novels published between ca. 1600 and 1830 has been established by the author. These properties include length, narrative perspective, generic subtitle, presence or absence of chapters and of inset narratives, subject matter, type of protagonists and several more. Based on these properties and their patterns of evolution, correlation and co-evolution over time, the author then creates a history of the French novel and its subgenres in the seventeenth and eighteenth centuries.

12.4 Annotating genre information within texts

Instead of generating various kinds of features describing entire novels, there are also some studies that investigate genre by adding and analysing genre-specific annotations to words or spans within texts.

An example of manual annotation of small spans within texts is Reinig and Rehbein (2019), in which the authors have manually annotated metaphorical expressions in a corpus of German expressionist poetry, that is a corpus which is homogeneous with respect to genre. Also regarding poetry, specifically Spanish-language sonnets, Navarro-Colorado (2017) presents a pipeline for adding annotations regarding syllabic structure of words and, building on this, metrical properties to individual verses.

Schöch et al. (2016) have first automatically tagged the full texts of a corpus of French novels with token-level semantic (WordNet) and morphosyntactic annotations. A sample of sentences were then manually annotated with respect to character vs. narrator speech. Based on these second-level annotations, a classifier was then trained to perform a sentence-level annotation of the entire corpus for the presence of character speech within the novels. They then use the information on the proportion of direct speech in each novel for investigations into genre-based differences. In a similar manner, Brunner et al. (2020) have used direct / indirect speech annotations to investigate the distinction between highbrow and lowbrow literature.

Kim, Pado, and Klinger (2017) have investigated genre in a subset of Project Gutenberg containing five subgenres of narrative fiction. They use sentence-based emotion annotation (Sentiment Analysis) for all texts and use this annotation for a genre classification task. In addition, they also conduct an investigation into emotional trends over the course of novels that form patterns

[1] See also Hettinger et al. (2016) for another perspecive of this data.

that are characteristic, to some degree, of the different subgenres. (See also Jannidis et al. (2017) for a similar investigation focusing on happy endings.)

Simple kinds of automatic, token-level annotations are often combined with automatic document-level feature generation. A study by Christof Schöch (2017b) combines token-level annotations and document-level feature generation. Schöch used topic modeling to automatically obtain thematic information about each play in his corpus as a way to investigate differences between comedies, tragedies and tragicomedies. As a part of the preprocessing for topic modeling, token-level annotation is often performed, for example by Schöch (2017b): both lemmatization, in order to focus on lemmata instead of word forms, and POS-tagging for the filtering of lemmata to include only content words such as nouns, verbs, adjectives and adverbs. Both steps reduce the dimensionality of the dataset and increase the semantic coherence and interpretability of the resulting topics. Generally speaking, this is done more often for languages other than English, as many of them are more highly inflected and lemmatisation has a larger impact on the results than for English.

One of the most elaborate examples of this strategy combining bibliographic metadata, additional document-level annotations relevant to genre and classification tasks based on most frequent words, is José Calvo Tello's study of the Spanish novel (2021), in which he has collected, primarily through close reading of the novels, a large number of document-level annotations regarding for instance the characters, themes, or the setting of the novels in his corpus, for subsequent use in a wide range of classification tasks. The author found, for example, that semantic features can be highly useful for subgenre classification of novels. In his case, these semantic features were obtained by annotating the tokens of the novels using a vocabulary derived from a semantically-organized dictionary, the *Diccionario Maria Moliner*.

12.5 Conclusion

As can be seen from this survey as a whole, the domain of computational analysis of literary genre has emerged as a dynamic area of investigations within CLS over the last 10-15 years. In terms of feature engineering and annotation, a wide range of kinds of features are now used to cover the many abstract properties relevant to literary genres and subgenres. It is striking, at least in this survey, that a majority of studies appear to focus on novels when investigating literary subgenres, with fewer studies on drama and even fewer on poetry. In addition, while in authorship attribution, complex processes of feature generation and annotation are rather the exception, the inverse is true for genre analysis. This may shift again, however, with the increasing use of deep learning for literary genre analysis, where active feature engineering as part of the preprocessing and annotation step may become obsolete, at the price, however, of decreased transparency and interpretability of the results.

References

See works cited and further reading for this chapter on Zotero.

Citation suggestion

Christof Schöch (2023): "Annotation for Genre Analysis". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological

Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/annotation-genre.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{schoch_2023_annotationgenre,
  title = {Annotation for Genre Analysis},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Schöch, Christof},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/annotation-genre.html},
  langid = {english}
}
```

13 Data Analysis for Genre

Julia Dudar (Trier)

13.1 Introduction

This text describes different methods of literary text analysis with a focus on literary genres. It is divided in four sections according to underlying methodology covering classification, distinctive features, clustering and genre-based corpus analysis.

13.2 Classification

This section describes studies that approach the issue of genre with a classification-based methodology. During the last years the automatic genre-based classification of long literary texts gained popularity among the researchers. This is not surprising, as this methodology gives an opportunity to investigate a high number of literary texts in a short period of time. Features that are used for classification can be extracted with different methods: most frequent words, distinctive words, stylometric analysis, or topic modeling methods. Moreover, such classification can be based on different algorithms: Naïve Bayes, SVM, logistic regression or K-nearest neighbours etc. There have been a considerable number of publications in this field, and we are going to introduce some of them trying to cover the variety of different approaches.

Underwood et al. (2013) define two challenges of automatic classification of literary texts according to their genres. Both of them are related to heterogeneity within works of one genre. The first challenge is a heterogeneity caused by the changes of texts across time, as literary history spans several centuries. The second challenge is the length of books: Literary text are much longer than journal articles and are internally heterogeneous. For this reason, they need to be segmented for classification. To address these challenges, the authors introduce a multi-layered solution with trained hidden Markov models, and several overlapping classifiers. For their analysis they use a collection of 469 texts from HathiTrust Digital Library. In their study they focus on relatively broad categories like prose fiction, nonfiction, drama, and nondramatic poetry.

Hettinger et al. (2015) classify a large set of German novels, experimenting with different machine learning algorithms and different types of features. In particular, they explore how different types of features affect the performance of different classifiers. Besides most common words and punctuation marks that they identify as stylometric features, they used topic-based features, extracted using an LDA algorithm, and features extracted from social network graphs (character and interaction graphs). In addition, a number of classifiers is implemented and evaluated for this task, among them Naïve Bayes (NB), Fuzzy Rule Learning (Rule), Multilayer Neural Network (NN) and linear Support Vector Machine (SVM). The data set consisted of almost 1700 novels either originally written in German or translated into German. From this data set domain experts identified 32 prototype novels, that belong either to a social or to an educational subgenre, after that 100 additional novels were labeled by exprerts as social or educational (labeled data). The classifiers and features were evaluated based on these two

datasets. The authors note that the combination of topic based (content) features and an SVM classifier yielded the best results.

The authors extended their work (Hettinger et al. 2016) by adding one further subgenre (adventure novels) and experimenting with feature engineering. As in the previous study, they used most frequent words (up 3000), character 4-grams, topic-based and social network features. However, they decided to concentrate on one classifier, namely SVM, as it showed the best results in previous analysis. They tried different combinations of features and experimented with the number of topics in LDA used for extracting topic-based features. The evaluation showed that classification results for adventure novels were much higher compared to the classification results of social and education novels. Network based features show the worst performance, while other content based features yield similarly high results.

Calvo Tello (2021) in his dissertation performed genre classification on different levels: fiction vs. non-fiction; narrative, drama or essay; subgenres of the novel. For the novel vs. non novel classification he used CORDE as training corpus, the largest existing historical corpus for Spanish, created by the Real Academia Española. He applied a grid search to find the best parameter combination for the classification of Spanish texts. He defined 20 combinations of features, including POS, tokens, lemmas, semantic annotation, most frequent tokens, mean and standard deviation of tokens etc. As classifier he chose logistic regression, as it yielded the best results for this task. The author also created the Corpus of Novels of the Spanish Silver Age (CoNSSA) and used it for subgenre classification task. He defined 11 subgenres, each of them was represented by at least 10 novels and applied multi-class classification using logistic regression. As the performance of classification was low, he transformed 11 classes into 3 more general classes: historical, comedy and naturalist. Alhough the performance in the second classification was higher than in the first one, it was still very low. Calvo Tello argued that a problem with the genre classification is that one particular literary work cannot only belong to just one subgenre, but a literary work is very often a mixture of different subgenres, which makes it really unique. This led him to the idea of multi-label classification, where each novel in the corpus was labeled as a binary vector, containing information to which subgenres the novel can be assigned. The results of multi-label classification were much higher above the baseline. The author also made the same observation as in the previous paper that some genres like adventure or erotic novels were classified more accurately compared to social and educational novels.

13.3 Distinctive Features

This section describes studies that approach the issue of genre with a methodology based on identifying distinctive features in a genre-comparative manner. The main idea of this method is based on the extraction of the most distinctive or characteristic words (known as keywords) from a target corpus in comparison to a more general and broad reference corpus. A target corpus usually used for this purpose consists of texts of one genre or subgenre, while a comparison corpus usually consists of texts of several genres and subgenres. Distinctive words are extracted with the help of so called keyness measure or measures of distinctiveness. There are several studies that are based on the application, analysis and comparison of different measures of distinctiveness (see e.g. Lijffijt et al. 2014; Paquot and Bestgen 2009), but there are only a few studies that are dedicated to the analysis of keyness measures used for genre comparison.

Schöch et al. (2018) describe in their paper an implementation of several variants of Zeta, used for genre comparative analysis. Based on the comparison of the three dramatic genres of comedy, tragedy and tragicomedy, they extracted distinctive words for these genres using Zeta. The aim of their paper was to reach a better understanding of the properties of Zeta as a measure for comparative analysis and to evaluate its usefulness for quantitative genre analysis. For example,

due to the analysis with Zeta, it becomes clear that tragicomedy and tragedy are much closer in terms of their vocabulary than tragicomedy and comedy, so that the tragicomedies can be best described as a special form of tragedy, not a special form of comedy.

Du et al. (2021) compared two dispersion-based measures of distinctiveness, namely Eta and Zeta, using a genre analysis task. In their study, they used a balanced corpus of 160 novels published in France between 1980 and 1989. 120 of them are lowbrow novels of three subgenres: sci-fi, crime fiction and sentimental novels. The rest 40 are highbrow novels. The genre analysis was based on a comparison of novels of one genre versus the three other genres. Their analysis was based on the distinctive words extracted by the comparison of novels. The authors of the paper came to the conclusion that both measures are able to detect meaningful and interpretable distinctive words for one genre compared to more general corpus of the novels.

13.4 Clustering

This section describes studies that approach the issue of genre with a methodology based on clustering, that is by creating data-driven groups of texts that are then related to genre-related metadata. Among the most popular clustering approaches used in the CLS Community are stylometry and topic modeling.

Coll Ardanuy and Sporleder (2014) clustered novels according to genres through building social networks. They collected a corpus of 238 prominent novels and presented the plot and structure of the novels in static (describes whole novel) and dynamic (describes one chapter) co-occurrence networks. After that the authors extracted a feature vector from each social network. They performed clustering over the obtained vectors, contrasting groups by genre and by author. For genre analysis they defined 11 most common genres.

Schöch (2017b) uses topic modeling to explore a corpus of French Drama of the Classical Age and the Enlightenment. The main goal of his paper is to discover the semantic types of topics that can be found in the collection of texts with topic modeling. Data-driven clustering of texts helps to investigate distinctive dominant topics and plot-related topic patterns in drama collections. The results of the analysis shows that different subgenres have their own lists of dominant topics.

13.5 Genre-based Corpus Analysis

This section describes studies that study literary phenomena not with a view specifically to genre, but using a genre or subgenre-based corpus. The phenomena analyzed are often genre-specific, however. For example, several studies focus on an analysis of different kinds of direct speech specifically in narrative texts such as novels and/or novellas (Brunner 2012, 2013; Schöch et al. 2016; Jannidis et al. 2018). Many studies perform network analysis specifically using corpora of dramatic texts (e.g. Powell 2014; Trilcke, Fischer, and Kampkaspar 2015), as the structure of interactions is often explicitly coded in dramas, in particularly when the texts have been encoded according to the guidelines of the Text Encoding Initiative. Also, many studies investigating the identification or classification of metaphors focus on poetry as their preferred genre (e.g. Tsvetkov et al. 2014; Shutova 2017; Do Dinh, Wieland, and Gurevych 2018; Reinig and Rehbein 2019). Of course, studies developing methods for metric analysis heavily focus on poetry (e.g. Hammond 2013; Carvalho, Loula, and Queiroz 2016; Navarro-Colorado 2017).

13.6 Conclusion

Summarizing all the information above we can conclude that there are three main methodologies used for literary text analysis with focus on literary genres: classification, clustering and distinctive features. The scope of such analyses is usually defined by text corpora that include literary works of several subgenres of one literary genre. Depending on the underlying methodology and particular implementation, data analysis can be based on most frequent words, content words or topic-words, gained through topic modeling approach, as well as other features. At this point it is important to mention that feature engineering plays a cruicial role for genre analysis. Besides these approaches there are also numerous studies that discover literary phenomena using genre-based corpora.

References

See works cited and further readings on Zotero.

Citation suggestion

Julia Dudar (2023): "Data Analysis for Genre". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/analysis-genre.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{dudar_2023_analysisgenre,
  title = {Data Analysis for Genre},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Dudar, Julia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/analysis-genre.html},
  langid = {english}
}
```

14 Evaluation in Genre Analysis

Julia Dudar (Trier)

14.1 Introduction

The issues relevant to evaluation in genre analysis vary depending on the methodological perspective adopted in a given study: classification, clustering, distinctive features or genre-based corpus analysis (see chapter "Data Analysis for Genre" (Chapter 13). However, there are some specific challenges for evaluation in terms of genre analysis. For example, the issue of hybrid genres, partial assignments or multiple assignments of texts can make evaluation more complicated. However, most researchers ignore this difficulty and just treat genres as categorical classes.

14.2 Classification

Some general remarks concerning classification approach can be found in the Chapter "General issues in Evaluation" (see Chapter 4). In classificatory approaches, and when relevant metadata is available in sufficient quality and amounts, evaluation can use standard statistical evaluation measures such as recall, precision, accuracy or F1-score. However when it comes to genre analysis, there are some specific challenges that should be considered in research design. The choice of the right evaluation measure is crucial and it depends on corpus design, its size and construction, and of course on the classification approach.

As mentioned in "Data Analysis for Genre"(see Chapter 13), Underwood (2014) classified a text collection from the HathiTrust Digital Library focusing on relatively broad categories like prose fiction, nonfiction, drama, and nondramatic verse. The authors experimented with different classification algorithms including random forests and support vector machines and also combinations of multiple algorithms. But their choice fell on an ensemble of regularized logistic models. It was trained by comparing each genre to all the other genres collectively. They choose this classifier not because of the best performance, but because it can be trained relatively quickly compared to SVM or other classifiers. With the high number of features and a range of different settings that should be tested, the authors decided that it would be the best option for them. For their task, SVM had slightly better performance but the implementation and training time were too long. A further advantage of regularized logistic regression is that it is highly interpretable. This characteristic gave the researchers the opportunity to find out unexpectedly important features of particular genres. Interesting is that most of this features were not content-oriented, but rather structural ones ("the", "my" for fiction genre). For drama the most weighted features were stage directions, past-tense verbs of speech but also structural features.

In their research on genre analysis, Hettinger et al. (2015) evaluated a variety of classifiers, including k-Nearest Neighbour (kNN), Naive Bayes (NB), Fuzzy Rule Learning (Rule), pruned and unpruned (Tree), Multilayer Neural Network (NN), and linear Support Vector Machine (SVM). The authors used a majority vote classifier (MV) as the baseline, which achieved an

accuracy of 0.66 for the prototype dataset and 0.58 for the labeled dataset. Based on the evaluation results, SVM performed the best for both the prototype and labeled datasets and for almost all feature types except social features. Naive Bayes also had high results, especially on the prototype dataset. Fuzzy Rule Learning performed the worst, with results below the baseline for all feature types on the labeled dataset. Regarding feature types, topic features showed better performance for all classifiers compared to stylometric and social features.

14.3 Clustering

In clustering-based approaches, evaluation is a little less straightforward, but measures such as Adjusted Rand Index or Silhouette coefficient can be used.

As discussed in the Chapter "Data Analysis for Genre" (see Chapter 13), Coll Ardanuy and Sporleder (2014) designed their genre analysis research based on the construction of social networks. On the basis of feature vectors extracted from social networks, they built clusters. According to their research design, the number of clusters was pre-defined and corresponded to the number of annotated classes (genres). Cluster evaluation was carried out with respect to the annotated data. However, the authors emphasize that this evaluation task was not trivial, as it was not always clear which labels correspond to which clusters. The label was assigned to clusters which contained most of the items of the labeled class. The authors used three metrics for the evaluation: purity, entropy and F1 measure. Coll Ardanuy and Sporleder did not use feature weights, however confessed that it could be useful, as some features had bigger impact on genre recognition than others. Despite of some weakness in research design (not consistent corpus design, no special treatment for multi-labeled texts etc.) the authors made some interesting observations. For example, they found similarities between historical, social and satirical genres: they all have a high proportion of minor or isolated nodes. On the contrary, *Bildungsromane* and picaresque novels often have one strong key protagonist and many minor characters around him or her. For science fiction, mystery and gothic novels, it is characteristic to have a mixed point of view.

As described in Chapter "Data Analysis for Genre" (see Chapter 13), Schöch (2017b) discovered dominant topics in French Drama applying topic modeling algorithms. While the authors of research mentioned above used pre-defined parameters in their analysis, the author decided to evaluate different parameters of topic modeling and to choose the best model for his research question. For this purpose he created 48 models based on a range of different settings, like varied number of topics or varying hyper-parameters. The evaluation of the models was based on a classification task, where the plays needed to be classified according to their subgenre. As input to this approach, the probabilities of each topic in each play were used. The author used four different classifiers (Support Vector Machines, k-Nearest Neighbors, Stochastic Gradient Descent and Decision Tree), while the performance of the algorithms was evaluated in a ten-fold-cross-validation setting. This classification task was solved with an accuracy from 0.70 to 0.87, with the highest results obtained by SVM. With the help of this evaluation, the author found the best parameters for his topic modeling model, which he applied on further experiments. The interpretation of topic lists showed that most of the topics have a high level of coherence and helped the author to discover distinctive topics of subgenres of French Classical Drama.

14.4 Distinctive Features

When using approaches based on the extraction of distinctive features, evaluation is particularly hard. There is no gold standard, and it is not possible to establish one. One alternative method

of evaluation is, again, to use downstream classification tasks.

Schöch et al. (2018) used classification for the evaluation of different parameters of distinctiveness measure. First, the authors performed keyness analysis with two variants of Zeta on two corpora: a collection of French Classical and Enlightenment Drama and a collection of Spanish-language novels from Spain and Latin America. To evaluate different parameters (segment size, number of segments) and variants of Zeta, they classified novels by genre (French collection) or by the continent of origin (Spanish collection) using distinctive words, as identified by the respective measure, as features. As classifier, a linear SVM was chosen. The result showed that most Zeta variants outperformed the baseline, while logarithmic Zeta had better performance compared to the Burrows Zeta. Segment size also influenced the performance: Burrows Zeta showed better results in the classification with large segments.

Du et al. (2021) used the same approach for the evaluation of 9 measures of distinctiveness, including two variants of Burrows Zeta. For their evaluation, the authors used a corpus of 320 contemporary French novels. This corpus contained the same number of novels for each of three subgenres of low brow novels: crime-fiction, sentimental novels, sci-fi and high brow novels. As in the previous study, the authors classified novels by genre, using top distinctive words, delivered by each measure, as features. But they expanded the previous study by adding seven distinctiveness measure, classifying the novels into four classes instead of two and testing the impact of the number of features on the classification performance. The most important result of their study is that measures based on dispersion or distribution (such as Zeta or tf-idf) are better suitable for the distinctiveness analysis compared to frequency based measures (such as chi-squared test or LLR test), as the former showed significantly better classification results than the latter, especially when using the smallest number of features.

14.5 Limitations

The issue of hybrid genres and multiple assignments of texts is still open, except some progress in this field made by José Calvo Tello in his dissertation (2021).

Hybrid genres can be seen as a mixture of different genres or as a new genre that combines features from existing genres. José Calvo Tello notes in his dissertation that in the evaluation of classification of hybrid genres, one of the main challenges is the lack of clear genre boundaries and the potential overlap of features. That is why it can be difficult to assign them to a single category (2021). One option to deal with this challenge may be the implementation of a probabilistic model that assigns multiple genre labels to each text. Another solution could be hierarchical classification system, where some genres are sub- or supergenres of other genres.

14.6 Conclusions

Overall, a research review on this topic has shown that evaluating of different approaches, like determining the most suitable annotation workflow (which could be machine-learning, manual or a combination of both) or selecting the best classifier, plays a critical role and has a significant impact on research results. While SVM is often a popular choice among classifiers in general classification research, applying a classification approach for genre analysis requires thorough investigation and comparison of multiple classifiers due to the unique feature properties and lack of clear boundaries between genres. The nature of literary genres is very complex, it means that literary genre can't always be assigned to a certain predefined category. That's why more

research in the field of hybrid genres and thorough investigation of genre characteristics is needed, before working on analysis and evaluation of quantitative methods in genre analysis.

References

See works cited and further reading on Zotero.

Citation suggestion

Julia Dudar (2023): "Evaluation in Genre Analysis". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/evaluation-genre.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{dudar_2023_evaluationgenre,
  title = {Evaluation in Genre Analysis},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Dudar, Julia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/evaluation-genre.html},
  langid = {english}
}
```

Part IV

Literary History

This part of the survey is devoted to issues of literary history.

15 What is Literary History?

Artjoms Šeļa (Kraków)

15.1 Introduction

Literary history, unlike other topics in this survey, is not a single task (like authorship attribution) or an object of study that could easily make up its own field in cultural studies (genre, gender, canonicity). It can enclose all its neighbors (history of genre, history of gender in literature), or be tangential to them, if history is understood as historiography: adding facts and documents to the body of knowledge. To be able to write about literary history, it must be first framed in one way or another.

15.2 Modes of literary history

Here, the focus is on the study of history in the sense of both Annales school and formalists. Fernand Braudel and colleagues saw the primary object of history in the study of change and continuity (see Le Goff 2015); formalists spoke of evolution of literature as an autonomous system, driven by forces that cannot be entirely reduced to social, political or economic currents (later this understanding was radically expanded by Bourdieu). In this light, the focus here is on the aspects of analyzing, inferring and evaluating *historical change* in literature — change that can be linguistic, morphological, or social. *Analysis* here mostly refers to detecting trends, trendlines and statistical models used in CLS to work with time. Inference points to methods of *anticipating* historical relationships between texts, often based on intertextual analysis: through techniques of text reuse, calculation of distances and building trees. The *evaluation* perspective describes attempts at historical generative models that try to replay history many times to see if baked-in assumptions about how things *may happen* can hold in sandboxes — radically simplified bare-bones worlds.

This perspective on literary history, of course, is quite specific and rooted in a certain vision of the main questions and methodologies. However, there are two modes of literary history that must be mentioned in this introduction, to better articulate what is not the primary focus of our survey (without excluding these perspectives completely).

The first mode, already mentioned above, is what can be called 'historiographic' — adding facts (Coker and Ozment 2019) and opening access to collections (Fischer et al. 2019; Schöch et al. 2021), contextualizing marginalized and forgotten authors (Wernimont and Flanders 2011; Borgo 2017), doing explorative dives into the Archive (Forlini, Hinrichs, and Moynihan 2016), encoding texts in a standardized way, preparing editions and critical apparatus (Mann 2018), or connecting various strands of knowledge together via the semantic web (Gehrke 2016; Schöch et al. 2022). In addition to that, we see certain work in this mode as 'reconstructive' , because it makes visible things that were not visible or accessible before: simulating paths of Romantic journeys through the Lake district using elevation data (Murrieta-Flores, Donaldson, and Gregory 2016), modeling early modern theaters (de Paepe 2014), or providing interactive

software environments to stage historical plays (Roberts-Smith, Desouza-Coelho, et al. 2013; Roberts-Smith, Dobson, et al. 2013).

At the most general level, literary history is simply a *story*, a *narrative*. This is the second mode, in which scholars recover, piece together, challenge, or reinforce the *stories* about literature. Computational methods is another way to gather evidence and argue about well-known problems: Black writers and Biblical discourse (So, Long, and Zhu 2018), representation of women novelists (Moravec and Chang 2021), pace of literary change and conflict for aesthetic dominance (Underwood 2019b, 2019a), the rapid diffusion of third-person narration or free indirect discourse (Long 2021), the changing fates of various historical forms of the novel (Paige 2020), or long-term continuity in usage of poetic meters (Šeļa, Plecháč, and Lassche 2022). This is a continuous and endless process of (re)writing the history, both at small and a large scales: it, of course, goes beyond any particular computational project, or specific text preservation / representation solution.

References

See works cited and further readings on Zotero.

Citation suggestion

Artjoms Šeļa (2023): "What is Literary History?". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/what-lithist.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{sela_2023_lithist,
  title = {What is Literary History?},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Šeļa, Artjoms},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/what-lithist.html},
  langid = {english}
}
```

16 Corpus Building for Literary History

Evgeniia Fileva (Trier)

16.1 Introduction

Corpus building in the field of literary history is a complex and challenging process. It is characterized primarily by the collection and modeling of large amounts of literary data (both metadata and full text) covering several time periods. These data are used by scholars to analyze language use, authorial style, genre changes, and other literary, linguistic, historical and social phenomena. In addition, corpora in the context of literary history ensure the preservation of valuable artifacts such as books, manuscripts, letters, etc. Collections of literary data offer for analysis an orderly and systematic amount of data that opens the way for scholars to study them thoroughly over time using computational methods.

Indeed, the digital age has brought about a wide range of resources, standards and tools for the construction of corpora and for collecting and processing information. As Merja Kytö (2011) notes, the turning point came in the 1970s and 1980s, when it became possible to piece together large amounts of textual information.

16.2 Early and/or large diachronic corpora

One of the first historical literary corpus projects is the *Thesaurus Linguae Graecae: A Digital Library of Greek Literature (TLG)*, which is a digital library that includes works of Greek literature from Homer to the fall of Byzantium in AD 1453 (and further into the modern era more recently). It was created in the late 1970s by Marianne McDonald and is currently housed at the University of California, Irvine. The TLG includes over 110 million words from more than 10,000 works by approximately 4,000 authors. It is designed to be a comprehensive resource for scholars, students, and anyone interested in Greek literature, language, and culture. Users can access the TLG through a subscription service that provides access to the TLG's database and search tools. The database includes a range of texts, from well-known works like Plato's dialogues and the works of Aristotle to lesser-known texts like the poetry of Hesiod and the works of the Church Fathers. Other pioneering, large and important digital libraries include the Perseus Digital Library, founded in 1987 by Gregory Crane, with a focus on Greek (32 million words) and Latin (16 million words) texts. Created even earlier, in 1971, by Michael S. Hart, the Project Gutenberg digital library remains an important source of texts, notably in English. The Oxford Text Archive (OTA) is another pioneering, large-scale text repository for the Humanities, founded by Lou Burnard and Susan Hockey in 1976.

Collections of historical literary data for the English language are traditionally the most elaborate and extensive. There are several large projects such as HathiTrust, Text Creation Partnership and Project Gutenberg to preserve and digitize large collections of text documents. For example, HathiTrust has about 17 million volumes in its digital library. From a historical perspective,

HathiTrust represents a significant achievement in the digitization of historical texts. By making millions of books and other materials available in digital form, HathiTrust has transformed the way scholars and researchers access and study historical texts. It has also provided new opportunities for the analysis and interpretation of historical data, offering new insights into the past and its relationship to the present. Metadata for HathiTrust are stored in MARC format and typically include information about the book uploaded to the library, such as author information, contributor information, publication information, contain identification numbers (ISBN, ISSN), etc.

More than 150 libraries around the world are participating in the TCP project to create reliable XML encoded ebooks. Such projects are very important to the scholarly community and provide the basis for many studies. The TCP Project uses SGML and TEI P3 (a subset of the TEI P3 schema) for text encoding. The original SGML DTD was replaced by an XML DTD for distributing and indexing the texts. The current standard for text encoding is TEI P5, but the TCP's XML schema is closer to TEI P4. Conversion between the TCP markup and TEI P5 is possible, and stylesheets are available to generate P5-conformant versions of TCP texts. These files are available for all TCP output, but in some cases, the P5 files may lag the TCP XML files in terms of revisions.

In addition, it is worth mentioning the Helsinki Corpus of English Texts. The Helsinki corpus is a collection of English texts spanning from the earliest Old English period to the end of the Early Modern English period. It contains about 1.6 million words and is divided into three main periods – Old English, Middle English, and Early Modern English - each with subperiods of 100 or 70 years. The corpus covers various genres, regional varieties, and sociolinguistics variables, including gender, age, education, and social class. Overall, the Helsinki corpus is a representative and comprehensive resource for studying the development of the English language over time. The Helsinki Corpus was updated (while preserving all information from the original version) using XML markup which ensures longevity and easy conversion to future formats. Valid and rules-compliant markup is crucial for automatic software, as even slight deviations can cause errors. The updated corpus follows the latest TEI guidelines, resulting in a need for considerable rethinking. The XML version is a single file with a teiHeader giving general data and individual text headers giving bibliographic and descriptive metadata. The annotation model used in the original corpus, COCOA format, was not fully convertible to XML using a single conversion script.

Other examples of relatively large, diachronic corpora of literary texts frequently used in CLS research include the TextGrid Digitale Bibliothek (containing a very large amount of texts in German, including translations into German, and covering multiple centuries); the Deutsches Textarchiv (DTA), intended as a diachronic reference corpus of German literary and non-literary texts (see Geyken and Gloning 2015) or Théâtre classique, a platform providing a large number of French dramatic texts, primarily for the time period ca. 1620 to 1810 (see Fièvre 2007). More recent players include DraCor, an innovative platform providing multilingual corpora of dramatic texts in multiple languages in an interactive environment based on the idea of 'programmable corpora' (see Fischer et al. 2019) and the European Literary Text Collection (ELTeC), a collection of corpora of novels in multiple European languages covering the period 1840 to 1920 (see Burnard, Odebrecht, and Schöch 2021; Schöch et al. 2021).[1]

[1]For more information on some of these corpora, see also the section "Corpus Building for Genre", Chapter 11.

16.3 The Use of Diacronic Corpora in DH/CLS Research

The historical corpora named above are essentially examples of diachronic corpora, which are characterized by the ability to trace the development and change of literature over time based on such a collection of texts. Diachronic corpora provide researchers with the opportunity to discover trends in the literature of a particular period and to analyze them (see also Chapter 18).

Diacronic corpora and historical texts collections underlie several of the studies we have found in our corpus of research literature. These studies have created research driven corpora, such as the one by José Calvo Tello, who has built the *Corpus of Novels of the Spanish Silver Age* (CoNSSA) based on texts obtained from a variety of sources, among them Wikisource, Gutenberg Project, Google Books, ePubLibre, Spanish National Library (BNE) etc. (Calvo Tello 2021). The time period covered by CoNSSA, however, 1880–1936, is relatively narrow for a diachronic corpus.

Based on the Project Gutenberg´s text collection, there is a study of "Quantitative patterns of stylistic influence in the evolution of literature" by James M. Hughes and colleagues (2012). This is an example of an examination of trends in authorial style over time. The researchers used a subset of works by authors from the Project Gutenberg database. The authors for a subcorpus were selected based on certain criteria (year of publication, availability of death and birth information about an author, at least five works presented in the Project Gutenberg collection), resulting in a final group of 537 authors. A representative feature vector was then created for each author by aggregating the frequencies of function words for each of their works, with a total of 7,733 works being analyzed.

Another interesting study was carried out by Trilcke et al. (Trilcke, Fischer, and Göbel 2016) on the basis of a popular corpus in German. They analyzed 465 German-language dramas from 1730-1930 for semantic connections in social interactions. The authors cite the Digitale Bibliothek in the TextGrid repository as the source for the corpus. The metadata was collected manually and stored in DLINA, a special XML format developed specifically for this study.

A study by Grace Muzny (Muzny, Algee-Hewitt, and Jurafsky 2017) proposes a new metric for measuring dialogism, that is, direct speech in novels over 230 years. The collection of texts taken for analysis includes 1,100 canonical English-language novels by 422 authors. The time period from 1782 to 2011, divided into 3 time periods (late 18th century, turn of the 19th and mid-20th centuries) is covered. In order to obtain a diverse corpus of dialogues, the authors developed new tools to extract dialogue from a large corpus of novels spanning three centuries, resulting in a new dialogue corpus with over 2 million instances of quoted speech. The corpus represents various genres and styles, and is crucial in understanding the abstract grammatical features that characterize spoken dialogue, which is essential for understanding literary style. It is notable that the researchers encountered a problem with the recognition of direct speech in the texts, because the OCR did not always recognize it correctly. For addressing this problem, they proposed the quote extraction system QuoteAnnotator.[2]

Sociological and cultural changes in literature are represented in the studies of Richard Jean So and Ted Underwood. Both authors explore gender in the context of literary change. So, Long, and Zhu (2018), for example, studied American novels from 1880-2000 and conducts a cultural analysis of racial and gender criticism. To do so, they analyzed some 10,000 books by 6,000 authors, manually annotated race and gender, and identified the 4,000 authors for whom this data was found. Based on this, the authors examine whether author style, language, and narrative depend on racial identification.

[2]Challenges in digitization of historical texts are presented in the chapter "Annotation for Literary History" (Chapter 17).

Ted Underwood has conducted a notable diachronic analysis of literature, namely on representation of women in literature. This study covers English-language fiction from the late 18th century to the early 21st (Underwood, Bamman, and Lee 2018).[3] Another study conducted by Underwood addresses the literary and historical development of genres. In his article on "The Life Cycle of Genres" (2017), he discusses the problem of historical comparison in literary studies and the difficulty in reaching a consensus about the life cycles of novelistic genres. Underwood cites Franco Moretti's research that suggests that genres display a rather regular changing of the guard with a 25-year rhythm. Similarly, Matthew Jockers' statistical model shows that genres framed on a 25- to 30-year scale are linguistically coherent phenomena in the 19th century. To investigate these questions, Underwood collected lists of titles assigned to a genre in 18 different sites of reception and gathered corresponding texts to compare groups of texts associated with different sites of reception and segments of the timeline to determine how stable different categories have been. The genres that Underwood used for his research include detective fiction, science fiction, and the Gothic. The author collected corresponding texts from the Chicago Text Lab and HathiTrust Digital Library to compare groups of texts associated with different sites of reception and segments of the timeline to investigate how stable different categories have been. The metadata were collectively developed at the Stanford Literary Lab and contain genre tags, dates and source description for a total of 962 texts.

Šeļa, Plecháč, and Lassche (2022) conducted a research study that provides evidence of the association between poetic meter and semantics in 18th- and 19th-century European literatures. The study uses five metrically annotated poetry collections in different languages (Czech, Dutch, English, German and Russian) to compare and analyze the metrical types used in the poems. The focus of the study is on iambic and trochaic metrical types, which are the most widespread ways to organize verse in European accentual-syllabic traditions. Data for the Czech collection is sourced from the Corpus of Czech Verse, the German data is from the Metricalizer corpus, and the Russian corpus is part of the Russian National Corpus. As for the texts in English, they are taken from the Gutenberg English Poetry Corpus, and the early modern Dutch songs are from the Dutch Song Database compiled and hosted by the Meertens Institute in Amsterdam. The focus of the study is on the Czech, German, and Russian corpora as they cover a comparable cultural niche and time span, while the English and Dutch collections are used as secondary sources to show the general validity of the study's claims for material with substantially different structures and origins.

16.4 Conclusion

To summarize, building a corpus for literary history is a challenging and multifaceted task that necessitates careful attention to several factors. It should be a representative selection of texts from various genres, authors, and regions that accurately reflect the literary period being studied. Once created, a literary corpus can be used to explore a range of research questions, from the study of individual authors and texts, to broader questions about literary movements, genres, and cultural trends. Through the use of computational methods, scholars can analyze large bodies of text in new and innovative ways, uncovering patterns and trends that might not be visible through traditional methods. Metadata information is not always present in studies based on diachronic corpora, but we still can observe that the presented metadata usually focuses on the year of publication and authors, since research based on diachronic corpora is viewed from a historical perspective. A sufficiently large and balanced corpus is necessary to ensure the validity of any results and avoid biases toward certain genres or periods.

[3] Readers can find out more about this study in the chapter on "Analysis for Gender" (Chapter 23).

References

See works cited and further reading for this chapter on Zotero.

Citation suggestion

Evegniia Fileva (2023): "Corpus Building for Literary History". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/corpus-lithist.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{filvea_2023_corpuslithist,
  title = {Corpus Building for Literary History},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/corpus-lithist.html},
  langid = {english}
}
```

17 Annotation for Literary History

Julia Dudar and Evgeniia Fileva (Trier)

17.1 Introduction

In the context of literary history, the annotation techniques and items that need to be annotated depend on the research question and the analysis methods used, such as clustering, keyness analysis, or classification etc. These methods have already been discussed in other chapters, see in particular "Data Analysis for Literary History" (Chapter 18). However, historical corpora raise some specific challenges for researchers. The issue of spelling variation and digitization quality are specific to literary analysis over time and undergo significant changes over time, something that can be particularly problematic for any text analysis method.

17.2 Digitization of Historical Print Media: Challenges and Limitations

Historical languages and old printings fonts were not invented for digital format, so the transfer from one medium to another is not without its difficulties. OCR or manual transcription inevitably results in errors, as well as entails interpretation (Piotrowski 2012).

In historical corpora, the language is also historical. In historical languages, there are often no standard variants of spelling and orthography, it is hard to determine the norm, and the spelling of the same word varies even within works by the same author. Correct spelling recognition is a critical step in the tokenization phase, so in the NLP preprocessing process, scholars are consulting lexical resources (such as WordNet), using informational retrieval and statistical methods from NLP (like POS tagging and lemmatization) to solve this problem (Piotrowski 2012).

The choice of digitization method is an important issue in preparing to scan a document. For handwritten texts, the method of digitization is chosen based on the condition of the document: problems such as damaged medium, bleed-through and fading especially complicate the task (Piotrowski 2012). Automatic handwritten text recognition (HTR) is a popular NLP research task today, and advances in recent years using neural network technology are impressive (see e.g. the Transkribus tool, frequently used in Digital Humanities, and a survey of its uses: Nockels et al. 2022). A tool gaining popularity in Digital Humanities for OCR of historical print is OCR4all (see Reul et al. 2019). However, in many contexts and depending on the materials, manual double keying by qualified personnel often still shows the most reliably accurate results.

The process of OCR also has a whole list of features to consider when scanning and processing the result. At the very beginning it is important to choose a suitable scanner (depending on the format and condition of the printed book or other materials) and the right settings. In addition, the quality of the text itself naturally affects the scanning result, and even at this stage obstacles can arise, since historical documents are often damaged, faded, or blurred, making it difficult

for OCR software to accurately recognize characters. Historical documents may also contain non-standard or archaic fonts that are not recognized by OCR software, resulting in errors and inaccuracies. OCR software may often have difficulty recognizing columns or lines of text that are broken or unevenly spaced, resulting in errors.

OCR technology has made significant progress in recent years, but there are still several limitations when it comes to historical texts. Researchers continue to explore new methods and techniques to improve OCR accuracy for historical documents. The creators of the open-source OCR tool OCR4all (Reul et al. 2019) offer researchers who work with historical texts the possibility to make correction during the process of text recognition, or even train own OCR model based on their corpora, if the gold standard is provided. This is especially useful for historical texts that may be difficult for traditional OCR tools to recognize due to factors such as aged or damaged paper, different font styles, or non-standard layouts. This tool enables the scholars to increase the performance of OCR recognition, to automate the digitization process, and to minimize the post-processing effort.

In some cases, the low quality of OCR recognition can be compensated for by using supplementary tools, manual correction or implementing more intricate annotation methods. For example, Muzny, Algee-Hewitt, and Jurafsky (2017) required a reliable dialogue extraction mechanism, to measure dialogism and analyze changes in the use of direct speech in drama over time. As the OCR quality of old texts was not sufficient for this purpose, the authors developed a tool for quote extraction and named it QuoteAnnotator. This tool is based on simple rules that enable the accurate extraction of dialogue sections from raw text, offering greater precision compared to regular expressions.

17.3 Normalization of Historical Texts

Spelling variations are a prevalent characteristic of historical corpora, as these texts typically have not undergone standardization. During the evolution of a language, lexical, grammatical, morphological, and syntactical changes naturally occur. Often these changes are well-known to researchers and do not pose significant difficulties in the preprocessing of historical texts. In cases where many changes occur, these historical properties can be considered as a separate language in their own right. To address this, special NLP tools already exist or can be created with requiring some additional effort. In cases where there are not much changes, an easy rule-based approach can be developed to address this task. For some research questions both language variants can be required: historical and modern one. In such a scenario, language modernization can be accomplished by adding an additional annotation layer while retaining the original historical spelling at the primary level (Bollmann 2018).

Further and more complicated problems for researchers are spelling variants that occur due dialectal influences or individual style of the writer. Even within works of one author, different spelling variations are possible. Such variations are often not known to the researchers and there are no special adapted NLP tools that could deal with such variants. As a result, time-consuming manual annotation often becomes necessary.

The issue of normalization of different spelling variants can be addressed through different solutions. They can broadly be divided into three categories:

- Domain adaptation: historical language is seeing as target domain, while the modern language is a source domain, the labeled data from the source domain are combined with labeled and unlabeled data from the target domain;
- Retraining the tool: an existing tool can be retrained using additional, manually annotated data (see Bollmann 2018);

- Data adaptation: historical variants are mapped to their contemporary equivalents (see Piotrowski 2012).

If the normalization is a necessary step of the preprocessing and annotation of the corpora, how can this step be automated? There are some approaches available to tackle this challenge. As mentioned earlier, some minor spelling variants can be solved using simple rule-based methods. Other approaches rely on the concept of similarity between historical and modern spellings and employ different string distance metrics to identify the closest equivalent of the historical word (Jurish 2010; Pettersson, Megyesi, and Tiedemann 2013).

In some cases, character-based statistical machine translation can be applied for normalization. In such scenarios, the normalization process is modeled as a machine translation of character sequences instead of word translation, as is usually the case (e.g. Sánchez-Martínez et al. 2013; Pettersson, Megyesi, and Tiedemann 2013; Schneider, Pettersson, and Percillier 2017).

Although neural networks are widely used for solving various NLP problems, their usage for spelling normalization is relatively infrequent. Inspired by machine translation techniques, Marcel Bollmann proposes a neural network approach for addressing historical spelling normalization tasks (Bollmann 2018).

17.4 Conclusion

Preprocessing and annotation of historical literary texts pose several challenges for researchers, including issues with spelling variations and the quality of digitization methods. OCR technology has improved significantly in recent years, but historical texts still present challenges due to damaged or faded documents, non-standard fonts, and unevenly spaced text. To address these issues, researchers sometimes need supplementary annotation tools. Concerning spelling variations there are different possibilities to address this challenge including domain adaptation, tools retraining, and data adaptation approaches.

References

See works cited and further readings on Zotero.

Citation suggestion

Evegniia Fileva, Julia Dudar (2023): "Annotation for Literary History". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/annotation-lithist.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{filvea_2023_annotationlithist,
  title = {Annotation for Literary History},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia and Dudar, Julia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
```

```
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/annotation-lithist.html},
  langid = {english}
}
```

18 Analysis in Literary History

Artjoms Šeļa (Kraków)

18.1 The Trend Line

If there is one method that unites diverse body of works in computational literary history, it must be a trend line. A plot with a line that is tracing change (or continuity) in some derived textual or bibliographical feature over time became the central device for making arguments and telling coherent stories about literary history at large and small scales. Decline of abstract lexicon in fiction (Heuser and Le-Khac 2012), continuity in genre classification strength over centuries (Underwood 2017), long-term stability of Ancient Greek literary style Storey and Mimno (2020), rise of dialogue share (Sobchuk 2016) and 'dialogism' scores (Muzny, Algee-Hewitt, and Jurafsky 2017), increase of linguistic repetition (Gemma, Glorieux, and Frédéric 2015), expansion and saturation of literary and publishing markets (e.g. Bode 2012), increase in character network complexity (Krautter 2020) and the fall of dramatic protagonist (Algee-Hewitt 2017), numerous keyword and topics trajectories (through one work, one author's lifetime, or large corpus). Different studies, relying on different information sources, text analysis procedures, historical and theoretical assumptions. What unites them is time which has a constant appearance as an X-axis on which different trends, bins, lines and curves unfold.

Inferring the trend in data — the average chronological direction of whatever variable currently sits on a Y-axis — in computational literary studies is routinely done with linear regression models that estimate the linear relationship between a response variable and time (y ~ x; Feature ~ Time). A regression line is then superimposed on empirical observations: an upwards slope will show an average linear rise of the Feature per unit of time, a downwards slope — average linear decline. When data follow clearly non-linear trajectories, researchers usually engage in curve-fitting: they use of exponential functions (Heuser and Le-Khac 2012) or polynomials of different degrees (Gemma, Glorieux, and Frédéric 2015; Trilcke et al. 2016); they rely on smoothing functions (often unspecified) that follow changes in data locally (Generalized Additive Models, Local regression Underwood and Sellers 2012; Underwood, Bamman, and Lee 2018; Krautter 2020; Pianzola, Acerbi, and Rebora 2020) or rolling means or medians (Erlin 2017; Sharmaa et al. 2020); they look at averages in a binned time over 2, 5, 10-year bins (Bode 2012; Sobchuk 2016) or eyeball empirical distribution of features through time.

What is central to the vast majority of observed studies is that linear models and smoothing functions are used as a simple trend-revealing technique, almost as a rhetorical ('see, this is *clearly* what is happening with this data in time'), not a statistical device. Indeed, literary history and causal inference are yet to stand together. With few exceptions, explicit statistical modeling that tries to tease out different factors that influence the distribution of a feature of interest, if it happens, happens *outside* of dimension of time. So et al. (So, Long, and Zhu 2018) use logistic regression to show a higher probability of Black writers using Bible citation in a social context; Koolen (2018) in her thesis models factors (gender, genre, being a translation) that might influence people judgement about literary quality; Manjavacas, Karsdorp, and Kestemont (2020) build bi-variate (outcome is coming from two predicted variables) model with group (or

'random') effects to estimate the relationship between quotation use in Patrologia Latina and surrounding topical context.

Exceptions include early work in literary sociology by van Rees and Vermunt (1996): authors engage with discrete-time event history models (also known as 'survival' models) to understand writer's debutes and factors that shape their reputation, measured by number of reviews. Essentially, they build a logistic regression for estimating a probability of increase in the number of reviews, but where each variable is allowed to change states at discrete events (i.e. publishing a new book). The advantage of using a survival model (which also pushes authors to bin count variables, like number of reviews, into discrete categories) over simpler multiple regression is not apparent from the study, but this research embraces complex modeling of change, unusual in the current CLS landscape.

More recently, Jockers (2013) and Underwood et al. (2022) treated time not as a neutral dimension that carries (and reveals) change, but as a factor of data. Jockers fit hundreds of independent linear regressions (one per predictor per textual feature) and tabulated statistically significant (based on p-values) results to estimate to which extent different factors — time, gender, genre, author — can describe observed similarities between texts. Underwood et al. (2022) used a similarly designed barrage of linear models (relying, instead, on the measure of explained variance, R^2) to argue that birth year, or generation, of authors, explains change in literary topics better than publication year, thus revealing a cohort effect in literary history. Additionally, their study asks how the use of topics change during individual careers. Do authors continuously 'update' topic use through their lifetimes, or use them somewhat consistently, according to some random baseline? Their research compares two processes with structural equation models that treat books as discrete events (similarly to Rees and Vermunt). On average, they find that topics that are driven by generation effects tend to reoccur in authors without a clear trend of usage, and topics that are resulting from specific historical times (war and other localized events) tend to be updated, revealing rising and falling patterns.

When it comes to detecting change in the usage of keywords or topics over time, Wadsworth, Vasseur, and Damby (2016) proposed a series of stand-out methods to trace the evolution of vocabulary. They focus on Sylvia Plath and look at cumulative distribution of each word through her works over time. The authors fit two simple functions to a cumulative trend — a linear and a power law. Using fits of two curves and differences between them, the study identifies groups of words with similar behavior: those that are used consistently over time, and those that are accelerating, or decelerating. Additionally, by looking at distribution of word-onset times (times before which Plath did not use a word regularly), the authors find clusters of words that only appear in specific periods and can potentially discriminate shifts in poet's style, or aid periodization.

We spent more time with these cases, because they present a more nuanced approach to chronology than simply fitting linear trends to data. The fascination with trend lines was already critiqued conceptually by Moretti and Sobchuk (2019): the authors note that trends smooth over complicated cases, conceal non-linearity and "remove conflict from history". Even if the line is smooth and monotonous, the forces that generated it might not be. Additionally, there are purely technical considerations. Out-of-the-shelf usage of linear models introduces the usual pitfalls known from other disciplines: over-reliance on significance testing (p-values) and explained variance (R^2) in the reporting instead of explicit statistical modeling; using linear regression, tied to the normal distribution, to model unsuitable data like discrete counts, probabilities and ratios, which can introduce unreasonable predictions — negative ratios, or probabilities larger than 1 (see Winter and Bürkner 2021 for a relevant discussion of using the Poisson distribution for modeling counts in linguistics).

A reliance on linear trend lines can also negatively impact the representation of uncertainty in the results when lines are fit on top of aggregated data: yearly averages and medians (e.g. Trilcke et al. 2016; Muzny, Algee-Hewitt, and Jurafsky 2017; Underwood and Sellers 2012; Šeļa and Sobchuk 2017). The resulting 'average of averages' line is fitted on a handful of yearly observations, conceal value dispersion and overemphasize trend direction in time. If the purpose of linear models is just demonstrating the trend, using empirical values together with indication of range of their distribution, or bootstrapped confidence intervals Sharmaa et al. (2020) might, in many cases, be a much more direct and transparent approach.

Overall, the analysis of chronological change and continuity in CLS is in the early stages of development. When compared to sophisticated text analysis, annotation and representation techniques, it is apparent that the dimension of time is a pretty much open methodological and conceptual problem.

18.2 Inferring historical relationships between texts

Chronological trends can be ubiquitous and useful, but they usually do not make any judgement about historical connections between data points, which is another large subfield of (computational) literary history. Which texts are connected through influence, citation and rewriting? Can similarity be used to also infer genealogy? How known manuscripts or editions of a text are related? These questions belong to the domain of intertextuality, and the main methodological domain of intertextuality, at least in CLS, is text reuse detection — techniques of tracing matching parts between texts.

18.2.1 Difference and similarity

Seo and Croft (2008) identify two major approaches to text reuse detection: string-based methods and similarity-based methods. While calculation of similarity (or distance) between full texts is rarely considered to specifically be a text reuse method, some studies (explicitly or implicitly) rely on language-based similarity to make arguments about influence and infer pathways of intertextuality. Jockers (2013) constructs networks based on pairwise stylistic similarities between texts and then identifies most influential texts as the most central nodes. A similar goal — of tracing patterns of stylistic similarity beyond authorship — is pursued by Eder (2017) in the so-called bootstrap consensus networks. Broadwell and Tangherlini (2017) use heatmaps based on distance matrices to understand boundaries of Modernism in Scandinavian literature — they use areas of overarching (dis-)similarity between books to argue about periodization. Iwata (2012) directly equates text similarity with historical relationships when dealing with collections of Japanese Noh plays. This equation is always a considerable leap of faith — there is no guarantee that similarity in frequency distributions of linguistic elements would also signal any kind of historical relationship. Stylistically central texts might be just the closest texts to an unobserved 'average' language — simultaneously similar to many, but related to none.

Recently, more attention was drawn to non-symmetrical measures of difference (Barron et al. 2018; Chang and DeDeo 2020) based on divergence. Divergence (specifically Kullback-Leibler divergence, KLD) measures the amount of 'surprise' of encountering a probability distribution P, given a prior probability distribution Q. Texts are usually represented as probability distributions over inferred topics (e.g. LDA) to keep divergence measures interpretable. Difference between $P \rightarrow Q$ and $Q \rightarrow P$ can, for example, signal an enclosement relationship, when one texts covers more ground than the other, topically focused and narrow. It is more surprising to encounter the general text having seen only the focused one, than the other way around.

It is argued that asymmetry in KLD is a good fit for modeling cultural point of views and their inherent subjectivity (Chang and DeDeo 2020). In relation to historical data, KLD was used to derive measures of novelty and resonance: the former describes how surprising is a text T given the past, the former describes how persistent is information from a text T in the future. High-novelty high-resonance texts introduce innovations that also leave their mark on the future. These measures were originally used to detect a novelty bias in the debates of the first revolutionary French Assembly (Barron et al. 2018) and recently extended to various historical cases, including the detection of events in Dutch chronicles (Lassche, Kostkan, and Nielbo 2022). The divergence-based framework provides an alternative approach to trends, focusing on uncovering relative, context-dependent positions of data points in time. The question of historical relationships between these points, however, remains open.

18.2.2 Text reuse and alignment

When two texts share matching or very similar fragments, it provides better ground for establishing (or suspecting) an intertextual link. The majority of text reuse research focuses on string matching: either for discovering potential quotations for future work and integration with search engines (Roe et al. 2016; Sturgeon 2017; Janicki, Kallio, and Sarv 2022), or deriving text reuse scores to understand the intensity of text reuse between documents (Bernstein, Gervais, and Lin 2015; Gawley and Diddams 2017; Shang and Underwood 2021).

There are numerous solutions and existing frameworks (like Tesserae Coffee et al. 2012; Tracer Büchler et al. 2014; Passim Smith et al. 2014; Text Matcher Reeve 2020) of text reuse detection. Most of them focus on matching literal, or near-literal, repetitions of strings that can be relatively easy solved by n-gram matching and sequence alignment algorithms that build up similar regions from local matching units. Many of the approaches originate in bioinformatics and are a part of the longest common substring problem (see Olsen, Horton, and Roe 2011). In literary studies, identified string matches are often weighted (e.g. by TF-IDF) to distinguish 'interesting' matches based on document-specific words from simple linguistic repetition (Bernstein, Gervais, and Lin 2015; Shang and Underwood 2021).

Tracing non-literal intertexts — like allusions — that are invoking the source text through a paraphrase, development of a shared theme or semantics is more difficult and rare (for an overview see Manjavacas, Long, and Kestemont 2019). Detection of allusions is an open interpretative practice: there is little ground-truth data available, because there is little scholarly agreement available, which complicates the adjustment of formal techniques to non-literal cases of text reuse. The nature of the problem that shifts the focus from lexical correspondence to semantic overlap, invites injecting semantic information into reuse detection.

Manjavacas, Long, and Kestemont (2019) – using a subset of manually identified Biblical allusions in writings of Bernard of Clairvaux – show that hybrid methods that combine both lexical and semantic information have the best performance in retrieving allusion sources. Specifically, they use the *soft cosine* that incorporates word similarity information directly into the cosine distance (which is calculated using bag-of-words vectors): the method can trace non-trivial similarities between target and source text.

Text reuse also poses a specific challenge to folklore studies that deal in large archives of written records of oral song performances. Automatic navigation through these archives is complicated because of high degree of textual variation, despite the large parts of the records are repeating on different levels. All texts, in principles, are related through a shared oral tradition, which conditions the reproduction of formulas and motifs, arrangement of parts and transmission of texts. Recent work by Janicki, Kallio, and Sarv (2022) shows that simple character bigram similarity of verse lines can be effectively used to identify 'equivalent verses' – clusters of very

similar strings. The study proceeds to align different texts in the collection using *cluster indices* instead of actual strings, which is aimed to generalize from strings to their *types* and serves as a good foundation for inferring similarity between folklore records with lots of variation. Poetic form makes the task somewhat simpler, since it provides a natural discrete 'unit' of text — a line, but this study is a promising development on the path of intertextual generalization.

18.2.3 Trees

A geneticist and one of the founders of cultural evolution, Cavalli-Sforza (with P. Menozzi and A. Piazza) wrote that "a tree can be viewed as a simplified description of a matrix of distances" (Cavalli-Sforza, Menozzi, and Piazza 1994, 33). Distances between anything: cities, species, languages, or individual texts. CLS routinely relies on dendrograms (that are built by grouping closest pairs in a distance matrix) as tools of unsupervised clustering. We are interested in a pattern of similarities and dissimilarities: which texts sit at the closest branches (maybe written by one author?), which texts form distinct clusters (maybe belonging to one genre?). These trees are not usually assuming any *historical* relationships between the leaves, which was the original purpose of a tree as a model: to sort out a history of life. Thus, the majority of CLS trees are not phylogenies and are of little interest to literary history.

Here we instead focus at the trees that *are* phylogenies and assume the descent with modification as a primary mechanism that produced its objects. There are at least three areas where phylogenetic methods or thinking is used in relation to (literary) texts:

1. Literary morphology. Gasparov (1996) reconstructed branching and merging history of European verse forms, while Moretti (2005) used a tree model to manually chart histories of clues in detective fiction and free indirect discourse.
2. Oral traditions: several groups of anthropologists use phylogenies to study branching of folktales (J. Tehrani, Nguyen, and Roos 2015; da Silva and Tehrani 2016) and mythological motives (Thuillard et al. 2018), an analysis that also opens connections to evidence coming from human migration history, paleogenetics and archaeology.
3. Paleography and codicology explicitly engages with phylogeny reconstructions: copies (or witnesses) of a manuscript are all related, but usually it is hard to tell how exactly: the historical record, similarly to archaeological one, is fragmented, incomplete; it needs hypothetical reconstruction. To do this, one can use accumulation of errors and changes in copies that, not unlike genetic sequences, reflects transmission history. Similarity between documents here *can* signal common origins. Use of phylogenetic trees in inferring relationships between manuscript copies is a modern extension of a stemmatology — manual philological reconstruction of manuscript histories (see recent handbook, edited by Roelli 2020).

Phylogenetic methods provide two clear advantages: automatic alignment (compare texts) and automatic inference of relationships (build a tree as a possible history). Phylogenies of texts started with relying on manual coding of traits: to test relationships between manuscripts of the *Canterbury Tales* (Barbrook et al. 1998), or reconstruct history of branching of *Little Red Riding Hood* (J. J. Tehrani 2013). Recently, sequence alignment is used more frequently for comparing strings of texts directly, as in the case of phylogeny of print editions of *The Wandering Jew's Chronicle* (Bergel, Howe, and Windram 2015).

Phylogenetic methods that originated in evolutionary biology obviously have design problems when used with cultural textual data: all 'ancestors' are thought to be unobservable (what we have is only 'leaves' of the tree) and often methods assume strong vertical transmission (like maximum parsimony trees, that rely on minimizing independent mutations). Manuscript

transmission simulated in the lab (Spencer et al. 2004), however, shows that these trees can still be useful and reflect historical clusters faithfully, but are not able to represent known ancestry relationships. Several methods outside of maximum parsimony trees were adopted that do not expect vertical transmission in data, e.g. split decomposition and, more recently, network-based approaches such as Neighbor-Net (Iwata 2012; Bergel, Howe, and Windram 2015). Few alternatives to phylogenetic framework as a whole were also proposed: Andrews and Macé (2012) model stemmata and witness readings as directed acyclic graphs, or DAGs, also used in causal inference — to utilize the formal approach, but escape the assumptions of evolutionary biology.

Since anthropologist Alfred Kroeber's critique in early 20th century, evolutionary trees are often dismissed as incompatible with with cultural histories where there is no reason to expect a strong tree-like signal: after all, culture mashes and mixes branches together. However, this is a misconception about evolution: branching, tree-like phylogenies are characteristic only for vertebrate evolutionary histories. Culture is more akin to viruses and bacteria with extensive horizontal exchange of information (and still, for various reasons, can be very tree-like, not fundamentally different from biological evolution (Durham 1990; Collard, Shennan, and Tehrani 2006). From adjacent disciplines we also see the expanding use of unrooted, network-based methods that do not assume a strong tree-like pattern in the transmission process: one such example can be the use of dynamic phylogenetic networks with community inference to trace histories of interacting individuals in electronic music (Youngblood, Baraghith, and Savage 2020). Patterns of collaborations produce distinct subgenres of music that stay recognizable despite an increase in transmission between branches in the age of internet.

18.3 Conclusion

We centered this survey on two areas of analysis that, we think, are indispensable for *historical* inquiry. First, usage of trends and trendlines deal with *describing* change: we see vast array of techniques and applications, but most research mainly stays at curve-fitting step and rarely engages in explicit causal inference. Second, distances, text reuse techniques and trees — all try to *infer* relationships between texts, often under historical assumptions that can be further used to reconstruct lineages of texts, traditions, or forms.

References

See works cited and further reading on Zotero.

Citation suggestion

Artjoms Šeļa (2023): "Analysis in Literary History". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/analysis-lithist.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{sela_2023_analysislithist,
  title = {Analysis in Literary History},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
```

```
  Survey Papers}} on {{Methodological Issues}})},
  author = {Šeļa, Artjoms},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/analysis-lithist.html},
  langid = {english}
}
```

19 Evaluation in Literary History

Artjoms Šeļa (Kraków)

19.1 Introduction

How can evaluation be framed in the context in literary history? All method-specific evaluation techniques for supervised classification (Underwood 2019a), clustering (Šeļa, Plecháč, and Lassche 2022) or information retrieval (Manjavacas, Long, and Kestemont 2019) are used in historical research. Often evaluation metrics are recycled to model a perspective, that is, to use them as a measure of label or class distinctiveness and trace the change in these values over time under different exposure of supervised model to parts of data. Are library-assigned 'detective' labels for books more recognizable than 'gothic' using only 19th century training data (Underwood 2019a)? Are future states of poetic traditions recognizable from the past (Šeļa, Plecháč, and Lassche 2022)? A so-called 'perspectival modeling' is not interested in maximizing model accuracy, but is using change (or, indeed, a continuity) of accuracy scores to argue about distinctiveness of periods (Broadwell and Tangherlini 2017), genre groups (Calvo Tello 2021), or pace of change in literary judgement (Underwood and Sellers 2016).

In this survey, I want to focus on an evaluation process that is specific to questions of history: the observed change and continuity itself. Statistical modeling against causal hypotheses is one way to approach this, but, as it follows from the chapter on "Analysis in Literary History" (Chapter 18), literary history is only starting to engage with causal inference. Another evaluative path is empirical validation of the observation, or robustness checks. These can come in many forms: by running replication analysis on corpora of different designs, by varying sampling strategies, establishing random baselines, performing bootstrapping or permutations.

19.2 A case in point: decrease of abstract lexical items

A case of a decrease of abstract lexical items in 18-19th century Anglophone fiction can be an informative example of how corpus selection and replication is coupled with empirical evaluation and evidence accumulation. Heuser and Le-Khac (2012) first find a *show, don't tell* effect in the development of fictional language: using the Chadwyck-Healey corpus (around 3,000 books at that time), they show a decline of abstract diction and increase in concrete words and action verbs. The authors link this to the change in the social space: novels stop being about private, inner worlds and open up to the public surroundings. Underwood and Sellers (2016), using another corpus of 4000 works assembled for the occasion, find a similar effect, but link it to a divergence of fictional language from non-fiction, which is further corroborated by Underwood (2019a) using vast HathiTrust data. In parallel, the same trends are observed when novels are stratified by canonicity (Algee-Hewitt et al. 2016), and, independently, by Kao and Jurafsky (2012) in their study of *Imagist language and influence in poetry*, where the authors assign the drop in abstractness to the specifics of the movement.

The observed trend is robust and ubiquitous. We know it is there, at least for English-language literature, but *why* is it there? There are multiple explanations: social change, diction divergence (emerging autonomy of literature?), literary movements. One of the problems with computational inquiry in literary history is that the answers we are seeking may very well not be in our data (or metadata). Luckily, there are some ways to deal with that, too.

19.3 Towards evaluation in literary history: generative models and historical simulations

Research in computational literary studies, and in the broader field of Digital Humanities, is mostly descriptive and data-driven. I use 'descriptive' here as an opposite to 'generative'. Scholars describe the *outcomes* of literary history: corpora, collections, trends, aligned texts, pathways of citations and try to understand the causes and drivers of observed patterns *a posteriori*. These claims might sound plausible, or counter-intuitive, but how we can tell which of those are more likely than others? The usual answer, in some parts of cognitive and social sciences is: controlled experiments. For complex systems and their macro-historical dynamics however, fully-fledged experiments are unrealistic (how would an experiment in evolution of dramatic structure look like? How could one in the rise and diversification of the European novel be designed?). In these cases, simulations and formal modeling is, often, the only way to directly engage with mechanics of history.

Simulations and generative models are, primarily, tools for understanding — and they are mostly ignored in CLS /DH , despite recently gaining traction in adjacent fields, like archaeology, that deal with similar problems of having an incomplete record of *outcomes* of cultural processes Acerbi, Mesoudi, and Smolla (2022). Formal models allow to explicitly define parts of the systems we are interested in and relationships between these parts (see Smaldino 2017 for a defense of formal modeling). The main advantage of simulations is that they put our verbal, informal models at risk, by forcing out our basic, deeply ingrained assumptions about important features of the objects and processes we study. In turn, the observed vs. assumed behavior of simulations allows better understanding and refinement of theories with their causality claims. In the last few decades, the modeling scene saw a rise of agent-based approaches to simulations (ABS): models that are not based on deterministic equations, but on the individual interactions between agents that can lead to complex, often not immediately intuitive, outcomes (examples of ABS implementations of classic models are sampling error or random drift in small populations, murmuration behavior in birds, or patterns of racial/social segregation).

There are only a few examples of formal modeling and agent-based simulations known to us that have been done for questions of literary history. Early attempts focused on simulating communication circuits in book history — an informal model of the book market by Darnton (Throne 2014). Gavin (2014), relying on Throne's model, in his essay tried to incorporate social simulations into the humanities. However, almost a decade later, a CLS INFRA survey (Van Rossum and Šeļa 2022) did note both nearly an absolute absence of simulations in current DH training and a lack of interest in the topic by practitioners. It suggests that formal modeling remains an alien, unknown approach with no immediately perceived usability.

Specifically important to literary history is a relatively unknown paper by Sack (2013) that directly challenges Moretti's claim that diversity of the novels in 19th century was driven by a rapid growth of reader's market. The model use authors as agents; authors try to figure out preferences of the readers (navigate the differently set up 'preference landscapes'); they do it by 'writing' novels (represented as binary strings); each new novel can be a recombination, mutation, or a direct copy of previous, relatively successful texts; under some conditions, there is also

a feedback loop between produced texts and readership market. Since simulated novels have explicit traits, their diversity can be measured and compared at the start and at the end of each simulated run of history. Simulations show that under the presented conditions, population size by itself does not guarantee growth in novel diversity (in case of homogeneous preference landscape diversity will actually decrease). Researchers reframe the problem as a general 'product diversification' process and suggest that an individual 'creativity factor' might be important for making novels explore diverse stylistic directions. This complicates the relationship between consumer population size, diversity and innovation and can drive further research towards the study of the very conditions that can *alter* innovative behavior itself (e.g. the competition between elites in the cultural field), which can be used to refine the model, which can be used to update the theory, etc, etc.

More recent examples of formal simulations for historical questions revolve around medieval manuscripts (paleography, thus, has all three components of a fully functioning scientific engine: empirical studies, experiments, and generative models). Kestemont and Karsdorp (2020) simulate the historical loss of manuscripts under very simple conditions to test their estimation of missing or undiscovered copies. They first simulate the very process of loss, where the 'true' population is known, and then, using the remains of simulated data, test how well the 'unseen species' estimator points to the the true value. This direction was recently radically expanded by Camps and Randon-Furling (2022) who simulate the whole process of manuscript writing, transmission and loss to not only argue about the discipline's assumptions, but also show the emergence of the frequent features of manuscript histories observed in stemmata reconstructions (e.g. the famous binary fork at the root).

Ferdinand Braudel, having a different kind of model in mind — the explanatory informal models of historical change — wrote: "Once the ship [the model] is built, what interests me is to launch it, to see if it floats, then to make it sail, as I wish, up and down the waters of time. A shipwreck always constitutes the most significant moment" (Braudel and Wallerstein 2009, (1958), p. 194-195). Simulations allow historians to launch thousands of ships, design shipwrecks and gain information that is hardly accessible otherwise. Current use of simulations in CLS demonstrates the potential of counterfactual literary history.

References

See works cited and further readings on Zotero.

Citation suggestion

Artjoms Šeļa (2023): "Evaluation in Literary History". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/evaluation-lithist.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{sela_2023_evaluationlithist,
  title = {Evaluation in Literary History},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Šeļa, Artjoms},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
```

```
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/evaluation-lithist.html},
  langid = {english}
}
```

Part V

Gender Analysis

This part of the survey is devoted to issues of gender in literary analysis.

20 What is Gender Analysis?

Evgeniia Fileva (Trier)

20.1 Gender studies as a field

Gender studies as a field investigates the ways in which gender is manifested in society, especially the cultural aspects that define one gender or another (Schößler 2010). Gender studies usually begins with a reference to the difference between gender and sex. Scholars agree on the following way of distinguishing these concepts, namely they believe that sex is a biological category that is by nature dimorphic, while gender is learned behavior that is influenced by social factors. In the context of the feminist approach, gender is not a dichotomous category, and gender boundaries are very blurred and difficult to define (Talbot 2019; Butler 1990). Talbot (2019) sees gender as a continuum in which there are different degrees of femininity and masculinity that apply to both genders. Butler (1990) believes that gender can be a "free-floating artifice". In CLS, however, the dichotomy of masculine/feminine is usually used because it is easier to implement, as Koolen (2018) argues.

20.2 Gender in literature

Gender studies is an interdisciplinary field that explores the impact of gender on various aspects of our lives. Although at first glance gender studies has much in common with sociology, this research field has a major role in literary studies. Literature from the perspective of gender studies examines authors, their personalities and authorial style, and characters, their behavior and speech markers that support or refute gender stereotypes.

Gender stereotypes is a problem which is mentioned in almost all the works we found on the topic of gender studies in the literature. Stereotypes are understood as social constructs of gendered behavior (Jockers and Kirilloff 2017), which scholars try to define in the context of a particular time period. For instance, in the Victorian period, women and men in society were assigned clearly defined social roles. While men were seen as proactive doers, women's position was primarily passive and submissive (Jockers and Kirilloff 2017). Such stereotypical gender traits are quite a strong conceptual construct, which has been formed and reinforced in society over long periods of time. Some stable stereotypes in the sphere of women's influence are such concepts as home, children, family, maternity, small spaces (kitchens, rooms in the house), as well as the whole sensual sphere, as opposed to "male" rationality and logic (Weidman and O'Sullivan 2017). These stereotypes led to the fact that typically female themes, which are mostly expressed in sentimental novels, were considered a "low genre" and did not become canonised. In addition, critics spoke quite negatively about women's writing, making no secret of their dismissive attitude. As a consequence, not only were women's novels published less frequently, but they were also less likely to win literary prizes, directly affecting the perception of women as writers (Koolen 2018). (On the relationship between gender, genre and canonicity, see also the chapter on "What is Canonicity and Prestige?", Chapter 25.)

One of the difficulties of gender-based research is that gender in the modern tradition is not a binary structure with clear boundaries. In the research literature, the concept of gender is closer to the biological concept of gender – that is, a binary structure where there are male authors and female authors (Jockers and Kirilloff 2017; Koolen 2018).

Another problem raised in many studies is the difficulty of determining the gender of the author, since there are uncertain cases. For instance, a woman could publish under a man's name (e.g. George Sand), a man wrote under a woman's name (e.g. Suzanne Vermeer), or it is difficult to understand the author's gender from a pseudonym. In these cases, stylometric methods of determining authorship and author gender are being implemented. There is also a case when the author is transgender (e.g. Maxim Februari), which makes it problematic to classify the author into one category or another (Koolen 2018).

20.3 Research and approaches

As a rule, the focus of gender studies in the literary context is style features and authorial markers, stylometry, comparison of gender representation (usually female) in literary works of different time periods, evolution of "women's literature," gender balance in literary corpus, etc.

Scholars extrapolate their findings more broadly to the social context and attempt to answer such questions as whether there is discrimination against female authors, novels and/or characters (Koolen 2018; Jockers and Kirilloff 2017) or what distinguishes female literary style from male (Rybicki 2015; Weidman and O'Sullivan 2017).

We cannot fail to mention Koolen (2018), who has done a seminal study of Dutch literature with a focus on the relationship between the author's gender and the perception of his or her literary work. She explores the so-called "prestige" and the path to literary acceptance, as well as its dependence on the author's gender. Koolen shows that gender and prestige or canonicity are closely related.

While the above-mentioned works relate to the author's gender, there are studies of gender tendencies in literary characters. For example, Jockers and Kirilloff (2017) have studied the relationship between the actions and personality of characters in 19th-century works and their gender. Underwood, Bamman, and Lee (2018) conducts a broader study of late 18th- and early 21st-century English-language literature. The focus of his research is on the conventional roles of characters that are dictated by their gender. He examines how much attention authors paid to female and male characters and what the distribution of roles looked like in several centuries of literature. In terms of the method of such research, parsing, namely recognizing characters in the text as well as the context in which they appear, plays a major role.

References

See works cited and further reading on Zotero.

Citation suggestion

Evgeniia Fileva (2023): "What is Gender Analysis?". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/what-gender.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{fileva_2023_gender,
  title = {What is Gender Analysis?},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/what-gender.html},
  langid = {english}
}
```

21 Corpus Building for Gender Analysis

Evgeniia Fileva (Trier)

Gender studies in literature that aim to analyze authorial style or character behavior are necessarily accompanied by a careful selection of material for the corpus. In the context of gender studies, the representativeness of the corpus is a critical factor in obtaining reliable results. In most of the papers that we found on this topic, the focus is on the number of male and female authors. For gender studies, it is desirable to build a balanced corpus, that is, one in which there is no overweight towards one or the other gender. Due to historical and social circumstances, as well as effects of gender on canonisation and canonisation, in turn, on preservation and digitization, almost all original corpora have more male than female authors. To make sure a comparison of novels based on author gender is possible without creating unsurmountable obstacles for corpus design, the creators of ELTeC, for example, set a very wide range for the proportion of novels written by female authors. Rybicki (2015) formed equal collections of texts by male and female authors for his analysis. The gender balance of the corpus is also addressed in studies by Koolen, Underwood, Calvo Tello, Schöch, and others.

21.1 Approaches to corpus building in gender-based research

Various approaches to constructing a corpus in the context of gender analysis were encountered in the articles from our corpus of research articles. We were able to distinguish two main approaches, namely the use of an already existing corpus, and the creation of a special collection of texts, taking into account the goals of the research.

21.1.1 Re-using existing corpora

The first type includes the work of Koolen (2018), who conducted a study based on the Riddle of Literary Quality corpus. This corpus was created by the Royal Netherlands Academy of Arts and Sciences (KNAW) to investigate the textual characteristics of Dutch contemporary fictional prose. Koolen herself was involved in the Riddle project and, in her monograph *Reading Beyond the Female*, has used the corpus created by the project for a specific task: the study of the relationship between gender and the (perceived) quality of literary texts based on readers' reactions to bestselling fiction. The original corpus is a collection of texts by Dutch authors from 2007-2012 and consists of 401 novels. Koolen provides a number of statistics on the gender distribution in the corpus, which is an important part of such studies. In the corpus mentioned, 55% of the works were written by men, 36% were written by women, and 9% have authors of unspecified gender (Koolen 2018).

Rybicki (2015) conducted a stylometric study using the Chawton House Library's corpus, which is a part of a digitization project at Chawton House aimed to build a historic collection of women's literature in English in the period 1600–1860. Rybicki compared word frequencies of the corpus with two reference corpora of famous female (22 novels) and male writers (21 novels) containing texts from the 19th and 20th centuries. The collection from the Chawton House corpus contained 46 novels written by women between 1723 and 1830. Interestingly, like the

Riddle corpus, this collection contains, among other works, text by anonymous authors that are very likely to be by women. In the course of the study, Rybicki (Rybicki 2015) combined materials from the reference corpora in different ways and used tool combinations for stylometric analysis.

Underwood, Bamman, and Lee (2018) completed one of the most extensive studies in terms of time period and number of units analyzed. They analyzed the gender development of characters in English-language literature over several centuries. They selected 104,000 books for their analysis and covered a time period of 306 years. The collection is based on books in English from the Hathi Trust Library, which have been compared with the Chicago Text Lab corpus and the Publishers Weekly collection to check for representativeness. The essay examines the changing importance of gender in fiction and characterisation from the late 18th century to the early 21st century. Underwood et al. argue that while gender divisions between characters have become less defined over time, there has been a decline in the proportion of fiction written by women, and the number of female characters has also decreased. Underwood's study is an example of the two approaches, namely author gender analysis and character gender analysis combined together, which is considered as an aspect that affects corpus composition. Underwood et al. emphasize the value of books drawn from academic libraries for their study because the significance of female characters there is greater and more transparent (Underwood, Bamman, and Lee 2018).

21.1.2 Creating customized corpora for analysis

The second type of approach to corpus building, where a corpus is created specifically for the research in question, includes the work of Weidman and O'Sullivan on the analysis of gender markers (Weidman and O'Sullivan 2017). They collected a corpus of 236 novels by 54 authors. Their analysis is based on an attempt to compare works in English from three literary periods – Victorian, modernist, and contemporary literature – in order to understand how an author's gender affects the lexical and conceptual apparatus of the work.

21.2 Concerns with gender in corpus building for other purposes

Researchers who do not work directly in the field of gender studies nevertheless often mention this aspect in connection with the corpus they use.

Thus, Calvo Tello (2021) also makes a comparative corpus analysis on the topic of gender balance. Using the example of the the *Manual de historia de la literatura española* (MdLE), he observes that female authors of Spanish literature in the time period 1880-1939 are much less frequently represented in this reference work than men: 6.5% versus 93.5%. Calvo Tello examines further the "importance" of authors present in this particular literary history. For his own corpus building work, he identified the statistical populations of the corpus based on certain traits and used the MdLE to define populations in a given time period. The novels used for defining the total population of (canonised) novels must be mentioned in the manual and meet certain criteria. He gathered the following information about the collected novels: author's name, birth and death year, gender, amount of dedicated pages in the manual and links to search for the author in digital libraries. His analysis shows that even though there are more men among the top authors, women's works are nevertheless just as important as those written by men, and are examined just as carefully (Calvo Tello 2021).

The *European Literary Text Collection* (ELTeC) pays a lot of attention to the gender ratio, which is one of the compositional criteria, in order to enable gender-based analyses. Thus, one of the

conditions for selecting material for the corpus was the presence of works written by women in the amount of at least 10% and at most 50% in each collection. The range of this percentage is so wide because in some languages and cultures, female authors are more strongly represented: for example, in English-language literature as opposed to Serbian, Slovenian or Czech literature (Schöch et al. 2021).

21.3 Genres and time periods in related works

Another aspect that is important in the creation and use of gender studies corpora is literary genres. Researchers have noted a correlation between the genre of literature and its perception by gender (Koolen 2018; Weidman and O'Sullivan 2017). For example, in the literary tradition there is the so-called "women's fiction" or "chick lit", which has a reputation as a low-prestige genre and is closely related to both gender and the personality of the author, as well as to the character and its behavior. This is due to the fact that the category of genre is not only a way of categorizing literature, but also a historical and social construct with its own stereotypes, tropes and patterns, which in turn influence the readers' perception of literature (Jockers and Kirilloff 2017).

We may be able to find further examples. In the Zeta project, it is obvious that there is very high proportion of male authors for detective fiction and science fiction, whereas there is a very high proportion of female authors for sentimental novels. However, a certain degree of bias through use of gender-specific, audience-related pseudonyms cannot be excluded.

It can be seen that in gender-based studies of literary corpora, the time period taken for analysis is a very important aspect. With the example of Underwood's study (Underwood, Bamman, and Lee 2018), one can see in retrospect how the importance of gender in literature has been transformed over a long period of time. The larger and more extensive the material for comparison, the better the overview of the importance of gender in relation to both authors and characters.

The study of a certain time period of literary history in a given language is also of great scientific and cultural value. Koolen gives an excellent overview of Danish literature (Koolen 2018), Calvo Tello takes a close look at literature in Spanish (Calvo Tello 2021). It is interesting to note that the Victorian period is often chosen for gender analysis in English-language literature. This is not coincidental, since during this period women authors became more visible to the literary community and tried to compete with men. Jockers and Kirilloff (2017) also note that female characters are changing and becoming more independent and retreating from the stereotypes dictated by the dominant position of men (Jockers and Kirilloff 2017). We can therefore conclude that the usefulness of Victorian literature is particularly great in the context of gender studies.

Thus, for gender-based corpus building used for literary studies, the balance of female and male authors is important, as well as the ability to recognize them automatically, which in the long run will be important for annotation and analysis.

21.4 Limitations

There are some challenges in building a corpus for gender studies. Firstly, metadata could be unavailable or incomplete. For example, while designing the corpora for ELTeC, it became clear that many library catalogues and other resources did not contain metadata about author gender and therefore, did not allow for targeted searches for novels written by women (Schöch et al. 2021). This fact also makes it difficult to identify anonymous texts or to identify the gender

of authors who who may have used a pseudonym. There are also cases where the author is non-binary or transgender (e.g. Maxim February), which requires the use of external databases to verify information about the authors (Koolen 2018). Sometimes, beyond subgenre, there are also correlations between author gender and other metadata categories, for example text length, as Schöch et al. (2021) indicate in the case of the Portuguese text collection. The lack of textual material and metadata, for female authors in particular, can greatly affect the results of analysis (Jockers and Kirilloff 2017).

Koolen (2018) discusses the importance of controlling for confounding factors in gender research when working with corpora by citing examples of studies that fail to account for potential biases, such as domain bias or publication bias. The author suggests that controlling for author and text type characteristics is necessary to avoid erroneously attributing differences to gender. Furthermore, the author notes that within the text type of fictional novels, there is a variety of subgenres that each have their own characteristics, which should not be attributed to gender. Overall, a corpus should be a sample that is statistically representative, and any findings derived from experiments using corpora are only reliable to the extent that this condition is fulfilled.

References

See works cited and further reading on Zotero.

Citation suggestion

Evgeniia Fileva (2023): "Corpus Building for Gender Analysis". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/corpus-gender.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{fileva_2023_corpusgender,
  title = {Corpus Building for Gender Analysis},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/corpus-gender.html},
  langid = {english}
}
```

22 Annotation for Gender Analysis

Evgeniia Fileva (Trier)

22.1 Main issues in annotation for gender

The methods of annotation are particularly well illustrated in the example of character gender research in the articles we have found. As for research on the author's gender, the annotation process is unfortunately not well-presented there. It can be connected with the fact that identifying and analyzing gender in literary texts is a challenging process, as gender markers are often not explicitly stated in the text or in bibliographic resources, especially for author gender. In contrast, characters' actions, dialogue, and descriptions are good material for studying gender markers, because there is a lot of textual material to be used. To address this issue, researchers have turned to annotation techniques, which involve labeling or marking specific linguistic features related to gender, such as pronouns, adjectives, or verbs.

Thus, syntactic parsing is a basic preprocessing step for research on gendered characters to understand what linguistic features they reveal. For example, Jockers and Kirilloff (2017) defined two classes of pronouns (male and female) as well as gendered nouns, and used the Stanford CoreNLP toolkit, in particular the dependency parser, for automatic parsing to identify pronoun-verb pairings in a corpus of novels. The output was processed to create CSV files containing the counts of verb associations with male and female pronouns. The data was then reshaped into a matrix with rows indicating source novel and pronoun gender, and columns indicating each verb associated with the novel and pronoun gender. Raw counts were converted to percentages to account for the imbalance in the occurrence of male and female pronouns. The dependency parser, however, was not 100% accurate, and the researchers conducted human analysis in order to find errors in the identification of verbs. To address this, the data was winnowed to only include highly frequent verbs and those that were most consistent with established patterns. The final list of 281 verbs was merged with book metadata to create a final matrix. The metadata included book information such as the file name, author's name, year of publication, and pronoun class.

Although gender recognition is a relatively unproblematic task, Underwood, Bamman, and Lee (2018) points out an aspect that can pose some difficulty, namely character name disambiguation. In the annotation process, it is important not only to identify a character's gender, but also to identify him or her by his or her own name. For example, Elizabeth Bennett may appear as Elizabeth, Miss Bennett, etc. In his work, Underwood uses the BookNLP pipeline, which is a Python library for analyzing literary texts.

Underwood's example of Miss Bennett points to another aspect important to annotation in the context of gender studies, namely the ambiguity of proper names in a literary text. In addition to different variants of the same name, characters may have statuses, professions, social roles, and nicknames that may indicate it in the text. In addition, it is not impossible that a single proper name can point to more than one character, as in the case of Miss Bennett. Disambiguation is solved, to some extent, by NLP tools.

Schumacher and Flüh (2020) discuss various annotation techniques for analyzing gender stereotypes and evaluations in 19th-century literature. They employ both quantitative and qualitative approaches, such as the analysis of pronoun usage and the examination of character traits and actions, to identify gendered patterns and dynamics in literary texts. As part of the _m*w_ project, Schumacher uses digital annotation, in combination with Named Entity Recognition and emotional analysis, as one of three approaches to gender role recognition. Character gender role identification involves three aspects – gender identification, character gender identity and character actions, and is determined by using personal pronouns, character qualities and descriptions of their actions. A model with preconditioned information about gender roles is run through the NER process, which results in the identification of categories such as "masculine," "feminine," and "gender-neutral." The corpus is annotated according to these categories using the CATMA tool, and subcategories corresponding to the character's social role (e.g. father, mother, child, etc.), gender identity (e.g. homosexual man), and personal qualities (e.g. narcissist) are also added. This way of annotation facilitates a careful analysis of the text for gender stereotypes.

To extract physical appearance features from the corpus, Koolen (2018) used Lexical Semantic Queries as one of three approaches (see the chapter "Analysis for Gender", Chapter 23) for more details). The Alpino parser was used as the method for developing queries to automatically parse Dutch novels and extract rich linguistic information, such as part-of-speech tags (including verbs and nouns) and grammatical functions (e.g., subject or object). The output of Alpino appears as linguistic parse trees in XML format, which can be queried using XPath. Koolen then constructed word lists of nouns and adjectives used in physical descriptions and included stative verbs in the queries. A set of thirteen XPath queries was developed based on the exploration of two novels and manual classification.

22.2 Limitations

Overall, these studies indicate that annotation for gender is a useful method for literary analysis that enables researchers to identify and examine gender-related linguistic features in literary texts. Nevertheless, it is necessary to take into account the challenges and limitations associated with this approach. For example, annotation by humans can be influenced by the subjective interpretation of linguistic features by different annotators. Furthermore, there is a risk of oversimplifying or essentializing gender by reducing it to a binary classification of male or female (Schumacher and Flüh 2020).

To address these limitations, scholars have proposed more nuanced and context-sensitive approaches to gender annotation. For example, Underwood, Bamman, and Lee (2018) argues for a performative approach to gender, which views gender as a fluid and contextual performance rather than a fixed identity. Rybicki's study (Rybicki 2015) proposes a statistical approach to gender annotation, using multivariate analysis of word frequencies to identify gendered linguistic features in literary texts.

In conclusion, gender annotation is a valuable tool for literary analysis that can help uncover the complexities of gender representation in relation to literary quality, authorship, and historical contexts. However, it is important to approach gender annotation with a critical and nuanced perspective, taking into account the fluidity and complexity of gender identities and avoiding oversimplification or essentialization.

References

See works cited and further reading on Zotero.

Citation suggestion

Evgeniia Fileva (2023): "Annotation for Gender Analysis". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/annotation-gender.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{fileva_2023_annotationgender,
  title = {Annotation for Gender Analysis},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Šeļa, Artjoms},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/annotation-gender.html},
  langid = {english}
}
```

23 Analysis for Gender

Evgeniia Fileva (Trier)

23.1 Introduction

Gender studies within CLS is a very active research area. The subject of study is gender, not primarily in the sense of biological sex, but as a set of social representations within the boundaries of certain socio-cultural perceptions that have been consolidated in a given society. According to this approach, gender is seen as an important concept of literature and appears as a dimension of social patterns of behavior that are rooted in a given type of culture. This approach can be observed in the scholarly publications we found. For example, Jan Rybicki or Sean Weidman and colleagues who study gender markers, Ted Underwood who focuses on the ways of gender identification of characters, or Corina Koolen who studies the influence of gender on the authorial style, etc.

What methodologies are used to address these issues? Koolen (2018) mentions two basic types of approaches: descriptive and predictive. The descriptive approach includes all techniques which serve to describe linguistic patterns and search for patterns in a dataset: tokenization, parsing, POS, etc. Descriptive approaches are often used in tasks such as sentiment analysis, named entity recognition, and information retrieval. As Koolen points out, this approach is easier to apply and organize the process. For further examples of the descriptive approach, one can refer, for example, to the studies such as Rybicki (2015) or Weidman and O'Sullivan (2017).

The predictive approach, as its name implies, serves to make predictions about data using machine learning techniques and models. It can include text categorization, machine learning model learning process, and evaluation of prediction results. This approach is expected to be more complex and demanding. This approach is demonstrated, for instance, by studies such as Schumacher and Flüh (2020), Jockers and Kirilloff (2017) and Underwood, Bamman, and Lee (2018).

However, as with stylometry, which can be applied both descriptively and predictively, many scholars combine both approaches. We have outlined the most significant studies in the analysis of gender in the literature that we found for this survey paper.

23.2 Current Applied Practice

Weidman and O'Sullivan (2017) pose the question of how gender affects literary style and whether male literary style can be distinguished from female literary style by gender markers. The authors use the distinctive words method and use a stylometric Zeta analysis and weighting Z-score differences. Using the stylo package for R (Eder, Rybicki, and Kestemont 2016), a comparative list of words that are most and least frequently used in the two corpora of novels by female and male authors was created. Two groups of markers, female (or 'preferred') and male (or 'avoided'), were involved in the Zeta analysis. A cluster analysis was then performed on the literature of three historical periods (Victorian era, modernists, and contemporary authors)

using Delta, which establishes text similarities based on most frequent words in the corpus. In this manner, authors who are considered more canonical and prestigious were clustered together, but another major distinguishing feature in the clusters was the individual authors' gender. The analysis showed the presence of stereotypical distinctive features, by which one can make a stylistic distinction between female and male authors. These features include the language of place, direction and location (e.g. "home", "kitchen" found in female authors vs. "country", "earth" etc. in male authors), the language of certainty (confidence for men, less confidence for women), preferred terminology (women have family and relationships, focus on interiority, while men focus on exteriority) and other clearly distinguishable linguistic markers. Although the trend of stereotypical markers in literature from all three periods persists, the authors note that contemporary women writers are increasingly less exposed to traditional stereotypical domains, which have evolved strongly over time.

A similar conclusion is obtained by Rybicki (2015), in which he examines gender-based authorial attribution in English fiction also using Zeta. On the basis of two corpora, 20th- and 21st-century and 18th- and 19th-century, a cluster analysis of Delta distances of most frequent words was carried out. Rybicki explains his choice of method and preference of stylometric tools over such methods as SVM by the fact that for literary-linguistic research, a more stable method should be used, even if there could be errors in accuracy.

The word frequency method was also used by Underwood, Bamman, and Lee (2018) in their paper on methods of gender identification of characters and analysis of the words used to describe characters in the period from the late 18th century to the 1960s. The authors work was based on a large collection of 104,000 texts taken from the HathiTrust corpus. A manually selected collection from the Chicago Text Lab (Chicago Novel Corpus) was used for comparative analysis. The BookNLP tool was used for gender identification, which showed good results especially among the descriptions of male and female characters. A manual sampling method was used to determine an author's gender, based on the Publishers Weekly list. After analyzing the statistical data, Underwood concluded that women were more likely to write about women and men, with men writing about men much less often. This trend has been stable over the entire time period taken. To examine how gender shapes character descriptions, Underwood used a representation that included different aspects of character simultaneously. A bag-of-words approach is used, where adjectives, verbs, etc. are employed to represent characters, while gender roles and personal names have been excluded. By labeling some characters with grammatical gender, the model can learn what words characterize "masculinity" or "femininity" of fictional characters based on the vocabulary associated with them. The accuracy of the model can show whether gender is a powerful organizing structure or whether it is becoming less prevalent. Even seemingly innocuous words can tacitly predict gender. The authors came to the conclusion that there have been significant changes in the representation of gender in literature over the last 170 years. The language used to describe fictional men and women has become less sharply marked, indicating that gender roles have become more flexible. Moreover, conventional binary roles have proven to be unstable over time, with shifting characteristics and attributes associated with each gender.

Unlike Underwood, who excluded descriptions of gender roles from his analysis, Schumacher and Flüh (2020) use them to train a model to be used in a tool for identifying gender in 19th-century German-language literature. First, 12 books by female authors and 12 randomly selected books by male authors were selected. The rest of the novels were used as a training corpus. Then all kinds of gender roles of characters (daughter, mother, husband, sexual orientation, social role, etc.) were integrated into the model. Three methods were used for analysis – Named Entity Recognition, stereotype annotation, and emotion analysis. Three gender categories were established for the Named Entity Recognition: male, female, and gender-neutral. The test was

conducted on several combinations of texts. NER results were added to the CATMA annotation tool and supplemented with sub-categories such as gender role names. With the help of CATMA, the authors also performed an analysis of emotions, identifying which emotions are most often mentioned in the test text. Further close reading and analysis of gender-specific emotions revealed several results, including, for example, that female characters more often show the emotion of fear and less often the emotion of anger than male characters.

Jockers and Kirilloff (2017) study the differences between female and male characters based on their actions, which they call "character agency". For this task, they use a classification method based on the nearest shrunken centroids classifier, which provides class predictions and probabilities, along with feature selection. This is especially relevant to this study, because in this way they can identify verbs that are important and not so important in gender identification and classification. The study found strong associations between verbs and pronoun gender in the corpus as a whole. The corpus was segmented by genre, and the model achieved varying accuracies in predicting pronoun gender, ranging from 58% to 100%, depending on the genre. The highest accuracy was observed in the anti-Jacobin, Evangelical, national tale, Gothic, industrial, and Newgate novel genres. The study suggests that the actions of characters play a crucial role in shaping our perception and comprehension of them.

In addition to descriptive methods, Koolen (2018) also uses a machine learning approach. Her research consisted of two main parts. In the first one, she tried to find out how gender affects an author's style. Three classification experiments were conducted using LIWC, machine learning and topic modeling, respectively. In addition to gender, variables such as country of origin and whether or not the author won a literary award were also taken into account. The LIWC method showed which groups of words were more likely to indicate female authors and which were more likely to indicate male authors. The experiment compared the two original corpora, namely the Riddle corpus and the Nominees corpus, through machine learning, trained the model using the Support Vector Classifier, then used a bag-of-words approach to calculate the relative frequency of lemmas, then performed an evaluation. Experiments showed that the machine learning method can determine the gender of the author of novels with an accuracy of 83%, with male authors being better classified than female authors.

The second part of Koolen's study focuses on the connection between physical appearance descriptions and gender. The first is to explore whether female authors really devote more attention to this aspect than men do. For this purpose she extracted descriptions of appearance from a corpus of chick lit novels. Three extraction methods were used – Lexical-Syntactic Queries, Machine Learning using SVM and a hybrid approach in which the result of the queries served as features in the SVM Classifier. As a result, Koolen concludes that manually constructed queries perform better than standard machine learning for extracting information about physical appearance in novels. Both methods have strengths and weaknesses, but the automated method is not robust enough for unseen text. A manual analysis of gathered sentences shows that physical appearance descriptions are abundant in literary novels and not necessarily more present in chick lit. Author gender is connected to differences in physical appearance descriptions, with male literary authors describing appearance the most.

23.3 Limitations and Ethical Issues

Gender studies in CLS reveals some ethical issues that must be taken into account in the analysis process. Koolen's discussion of this topic is very well presented in her book *Reading beyond the female* (2018). There are unavoidable problems with gender categorization, as gender is a primary characteristic that people use for classification and can lead to stereotyping and essentialism. The use of NLP and big data can exacerbate these issues, making it a pressing

topic for the field. Koolen discusses also the issue of neutrality and objectivity in NLP techniques, particularly in machine learning models that can identify stylistic differences between male and female authors. However, researcher's selection of gender as two groups for training the algorithm means that the technique is not neutral and has similar issues to the descriptive method. The predictive success does not necessarily reflect the explanatory value of the gender division. Thus, NLP needs to keep the same issues in mind as the descriptive method.

Koolen discusses also the potential for bias in gender research using corpora. One issue is the violation of the assumption that a corpus is a statistically representative sample. Confounding factors such as the types of products that men and women tend to write, read or review can affect the validity of conclusions drawn from the dataset. Another issue is the potential for publication bias. Controlling for author and text type characteristics is important in gender research using corpora. Even within the text type of fictional novels, subgenres have their own characteristics that might be erroneously attributed to gender. The second issue argued by Koolen is that researchers often accept gender as a cause of difference without seeking supporting research beyond the chosen dataset, which can lead to bias and promote the separation of female and male authors in literary judgments. As a good example of acceptable gender-related text research, Koolen highlights Rybicki (2015), who examined whether a corpus supposedly consisting entirely of female authors might actually include works by male authors.

Limitations connected with the corpus have also been noted in some other works we have mentioned. For example, Weidman and O'Sullivan (2017) argue that their conclusions are only true for their dataset and leave out some aspects, such as the collective macro contribution of women to the literature. Jockers and Kirilloff (2017) note that their results might have been different if they had a larger corpus and richer metadata. In addition, their results support gender stereotypes to some extent. The potential of such research lies in the consideration of irony, a wide range of forms of character presence (e.g. first-person forms), and narrative time.

References

See works cited and further readings on Zotero.

Citation suggestion

Evgeniia Fileva (2023): "Analysis for Gender". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/analysis-gender.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{fileva_2023_analysisgender,
  title = {Analysis for Gender},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
```

```
  url = {https://methods.clsinfra.io/analysis-gender.html},
  langid = {english}
}
```

24 Evaluation for Gender Analysis

Evgeniia Fileva (Trier)

24.1 Introduction

The evaluation of gender analysis in CLS involves assessing the selected methods and their effectiveness for classification and prediction tasks. Despite the wide range of tools and techniques used for gender research in literature, not all of them demonstrate high efficiency. There is an active discussion in the CLS and DH community on this topic. Rybicki (2015) argues that there is no universal consensus on the optimal method, and comparative studies only show slight improvements. Based on this, he concludes that the choice between different methods is not significant in literary studies, and stable methodology is more important. The main task for research in this area is to search for visible gender signals and their analytical analysis, as well as predicting gender on unseen material, both for characters and authors. The following text highlights research in which the authors provided an evaluation of their methods and/or results.

24.2 Evaluation of gender signals

One of the methods for identifying the gender of characters in literature is BookNLP, used by Underwood, Bamman, and Lee (2018). BookNLP has shown a high level of accuracy in assigning genders to characters based on their names and honorifics, with a precision of 94.7% for women and 91.3% for men. However, there were difficulties, such as recognizing the gender of first-person narrators in cases where there are not enough gender-specific references to the pronoun "I". Regarding the predictability of grammatical gender from ungendered evidence using characters from novels in HathiTrust, Underwood, Bamman, and Lee (2018) have found that the accuracy of gender identification decreases from 1840 to 2007, and this trend is consistent across different sources of data and modeling strategies.

When comparing characters created by women and men within fiction written by either gender, the models are consistently less accurate for characters created by women (by 2.5% on average). Gender differences seem to be more pronounced in stories written by men. The accuracy of models that attempt to distinguish character gender in groups of characters drawn only from books by men or by women varies by around 10% across 230 years, from roughly 76% to 66%. Underwood points out that it is unclear whether this constitutes a dramatic or a subtle change, and the strength of 76% accuracy is uncertain given that the model has only 54 words of evidence, on average, for each character. Biographical information about the author is hard to infer from this limited data.

The method of classification based on the verb-pronoun word combination and the study of external factors affecting the classifier's ability to predict pronoun gender based on verb associations, conducted by Matt Jockers and Gabi Kiriloff (2017), showed the following results. 10-fold cross-validation was performed, and an overall accuracy of 81% was observed, with an error rate

of 16% for male pronouns and 22% for female pronouns. The authors also conducted hold-out validation, which showed a 30% improvement over chance, suggesting a strong association between certain verbs and pronoun gender. The verbs "wept," "sat," and "felt" were associated with female pronouns, while "took," "walked," and "rode" were associated with male pronouns. During the analysis, ten verbs (five male and five female) that were most useful in differentiating between male and female pronouns were identified. The study found that female pronouns were slightly less gendered or "codified" than male pronouns, as some verbs typically associated with female pronouns were still used with some male pronouns. The study also found that the algorithm was less confident in its assertions about verbs associated with female pronouns, suggesting that these verbs are generally more ambiguous in projecting a clear pronoun gender class. Overall, the algorithm reported an 81% accuracy rate in predicting pronoun gender. The researchers then segmented the corpus into different genres and found that the overall prediction accuracies were sustained, with accuracy ranging from 58% to 100% depending on the genre. The highest accuracies were observed in the anti-Jacobin, Evangelical, national tale, Gothic, industrial, and Newgate novels. While author gender was not found to be a strong determiner of classification accuracy, there were differences in how male and female authors associated male and female pronouns with verbs. Male authors were more likely to create female characters that defy gender stereotypes, while female authors were more conventional when creating characters of their own gender. When author gender was unknown, the machine struggled more with predictions of female pronouns.

Corina Koolen's study (Koolen 2018) compared the performance of lexical-syntactic queries, an SVM classifier, and a hybrid method for identifying sentences containing descriptions of physical appearance. The evaluation was done using precision, recall, and f-measure, and an unweighted average was used to account for the small percentage of sentences containing physical descriptions. A macro evaluation that averages scores on both classes yielded an f-score of 30% for sentences containing physical descriptions and 90% for those that did not, but only the f-score for the former was reported in the study to avoid unjustly inflating the outcome.

Several elements were fine-tuned to improve their performance. These elements include adjusting and adding queries, machine learning features, and testing the size of the lexicon. However, none of these elements improved the overall outcome, and cutting the original lexicon in half caused a 5% drop in performance. Therefore, Koolen concluded that enlarging the lexicon might be the easiest way to improve performance. The hybrid method outperforms the SVM classifier, with the former being particularly effective in classifying chick lit compared to literary novels. Chick lit features more varied descriptions of physical appearance than literary novels do, with all degrees of comparison of the word "beautiful" appearing as a discriminating feature. Nonetheless, Koolen notes that one should be careful in interpreting machine learning features.

24.3 Conclusion

While the methods described here of analyzing gender in literature have shown varying degrees of success, they also face challenges such as ambiguous pronoun references, changing language use over time, and author gender biases. Fine-tuning elements can improve performance, for example, one of the ways to improve outcome could be enlarging the lexicon. The fact that some studies observe a change in performance depending on literary period, with accuracy dropping over time, shows that performing such evaluations in CLS research is not only a best practice that is essential for an assessment of one's methods and the degree of trustworthyness of one's results, but can also provide insight into the history of gendered writing in its own right.

References

See works cited and further readings on Zotero.

Citation suggestion

Evgeniia Fileva (2023): "Evaluation for Gender Analysis". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/evaluation-gender.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{fileva_2023_evaluationgender,
  title = {Evaluation for Gender Analysis},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Fileva, Evgeniia},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/evaluation-gender.html},
  langid = {english}
}
```

Part VI

Canonicity and Prestige

This part of the survey is devoted to issues of canonicity and prestige.

25 What is Canonicity?

Lisanne van Rossum, Karina van Dalen-Oskam (Amsterdam)

25.1 Introduction: canon and canonicity

The word 'canon' goes back to the biblical canon, a collection of scriptures that solidify authoritative doctrine in Christianity. Analogous to the biblical canon the literary canon is generally understood as the traditional curriculum of literary texts (Guillory 1993, 6), or in somewhat more detail "a constellation of highly valued, high-cultural texts that have traditionally acted as arbiters of literary value, determining the discipline of literary studies as well as influencing the critical and cultural reception of literature." (Mukherjee 2017). The literary canon is also used more broadly as a reference to works with literary connotation (Harris 1991, 110). Key is that the canon is a selection, a small sub-set of everything published, and reflects the literary prestige the texts have in the eyes of readers, critics, publishers, fellow-writers, and teachers. It is important to keep in mind that relative cultural meanings of both canonicity and prestige exist – and flourish – in all kinds of textual spaces, for example in fan fiction (Chin 2018, 250).

Harris posited that the literary canon may shift in its exact composition and position, but that the purposes of the literary canon are distinct and worth further investigation (Harris 1991, 110). Forty years down the road, national literary fields now call for renewed evaluation of the contemporary relevance and validity of the hegemonic canon, with special focus on whose interests the canon should reflect. The canon is selective and inherently defined both by what it contains and what it excludes. The current literary canon, Karina van Dalen-Oskam observes, is the product of systemic prescription and re-inscription of prestige (van Dalen-Oskam 2021). Contemporary debates are centred on the democratization of the canon such as collective reading initiatives in the United Kingdom;[1] on reflection on a culturally descriptive canon through a reader survey in The Netherlands (Koolen et al. 2020; Van Deinsen, Sevenants, and Van de Velde 2022); on the formation of the nation state through politicized canonization efforts in Belgium (Tollebeek, Boone, and Van Nieuwenhuyse 2022); and on the potential of the canon to bring minority perspectives into view, such as in Germany (Meyer 2021).

25.2 Analysing canonised and non-canonised texts

Textual analysis of the literary canon has fundamentally changed since the digital turn. In practice this means that researchers no longer need to limit their analysis to a small selection of literary texts. Pioneering work was done by Matthew Jockers in his analysis of a corpus of 19th-century English-language novels (Jockers 2013) and at the Stanford Literary Lab, which dedicated at least three of their influential Pamphlets to large-scale analysis of literary text corpora, considering both canonicity and prestige in their approach (Algee-Hewitt and McGurl 2015; Algee-Hewitt et al. 2016; Porter 2018). Jodie Archer and Matthew Jockers gathered a corpus of novels on the New York Times' Best-seller list to explore the differences between these

[1] See e.g. Libraries Connected's The Novels that Shaped Our World programme

novels and those that did not reach the same level of popularity, also taking into account different levels of literary prestige (Archer and Jockers 2016). Mass digitization of library holdings offers the opportunity to make larger and different selections of texts and an enlarged repertoire of options to study them, as Ted Underwood stated. "Instead of arguing about samples as if they were competing canons, we can adopt a relational mode of reasoning about literary history, akin to the methods of social science" (Underwood 2019a, 177).

Social constructivist theories of prestige as cultural capital revolved around the question of who or what creates prestige, focusing not on the art works themselves but on sociological processes. In his work The Field of Cultural Production, Bourdieu outlines three sources of distinction, or "three competing principles of legitimacy": artists' in-group approval, 'consecration' by the institutions of the dominant elite, and a moderate amount of popular success (Bourdieu 1996, 50–51). In turn, Gramsci countered that artistic approval is much more slippery than Bourdieu presented, as the peer group is relative to each author and their intentions (Gramsci 1991). Guillory argued that 'the school' in the sense of the Academy is the main factor in creation of cultural capital and canonicity (Guillory 1993). Despite these expansions on Bourdieu's conceptualization of distinction as a diffuse market of exchange between forces and actors in the literary field, the art work itself was not systematically inserted back into models of prestige until De la Fuente and others proposed the New Sociology of Art (De la Fuente 2007).

In line with this New Sociology, recent academic inquiry into perceptions of literariness has shifted towards a more integrated approach of both textual analysis and large-scale study of literary value judgments in the project *The Riddle of Literary Quality* (Koolen et al. 2020; Van Dalen-Oskam 2023). Modelling prestige in this way can give way to quantitative discussion of those textual properties that are associated with prestige within different temporal and spatial contexts, as exemplified by Van Cranenburgh's work on literary language in the Dutch market (van Cranenburgh 2016). As suggested by Ted Underwood, consciously modelling with rather than against the bias of a corpus can also present new comparative opportunities for Digital Humanities research (Underwood and So 2021). Acknowledging the historically specific nature of a language model, for example, allows us to evaluate how a seminal novel from the past would have performed today. Indexing canonicity over time and place can trace developments and changes in cultural value placements, an approach taken by Royal Academy of Dutch Language and Literature's Canon Festival and the sister projects The Riddle of Literary Quality by the Huygens Institute for History and Culture of the Netherlands and the Novel Perceptions: Towards an Inclusive Canon project from the University of Wolverhampton.

Another approach to canonicity combines an analysis of the language used in the texts' reception, such as commercial reviews, academic papers and publications, and public Tweets or Goodreads reviews, with that of the texts themselves, as investigated by the Impact and Fiction project by the Huygens Institute for History and Culture of the Netherlands.

25.3 Conclusion

The literary canon is formative in all stages of cultural production and reception: it contributes substantially to defining what is written, published, read, reviewed, appreciated, and archived. As such, understanding the canon will lead to increased understanding of those texts that are available for computational analysis, either in digitized source material or as a node in a network of sources surrounding a text.

References

See works cited and further readings on Zotero.

Citation suggestion

Lisanne van Rossum, Karina van Dalen-Oskam (2023): "What is Canonicity?". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/what-canon.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{vanrossum_2023_canonicity,
  title = {What is Canonicity?},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {van Rossum, Lisanne and van Dalen-Oskam, Karina},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/what-canon.html},
  langid = {english}
}
```

26 Corpus Building for Canonicity

Lisanne van Rossum (Amsterdam)

26.1 The corpus and the canon

Data sets define the object of inquiry. Corpus building is therefore an important methodological step for the textual analysis of canonicity and literary prestige. The canon and the academy are themselves intrinsically connected: a place in the canon facilitates the availability and transmissibility of a text and this availability in turn aids its institutionalization in research, an idea that is much conceptualized in academia (Moretti 2000; Van Rees 1983, 1987; van Rees and Vermunt 1996; Bode 2020). As computational rendering of a complex sociological process of value attribution, the corpus thus is both the result of canonicity and responsible for canonicity.

A methodological overlap between corpus building and canonicity is that they are limited to what they contain and defined in large part by what they exclude. To paraphrase: selection is central both to canonicity and to corpus building. In computational literary studies, such boundaries can be approached as discrete by using pre-selected lists. Marc Verboord writes that "…a common way of dealing with the problem of selection, besides ignoring it, is to refer to selections other actors in the literary field have already produced" (Verboord 2003, 260).

However, working with data is frequently based on arbitrary and less evident criteria, especially in the case of prestige. Verboord poses that a main challenge on the methodological level "is that the term 'canon' suggests that literary quality is dichotomous in nature: an author either belongs to the 'canon' or s/he doesn't. In reality, more levels can be perceived. As will be shown, these levels are largely a result of the various dimensions of which literary prestige consists" (Verboord 2003, 261). The formation of a corpus to operationalize canonicity, as such, presents an act of approximation. In the end, the results produced through corpus building remain relative to the data set and careful consideration of corpus composition results in a more robust understanding of its outcomes.

26.2 Current applied practice

Recent studies have frequently taken the route of using one or more actors or indicators in the literary valuation process to build a corpus for the analysis of prestige. The Stanford Literary Lab, for example, has focused on expert opinions as the basis for their canonicity analysis by using the *MLA International Bibliography* as a corpus to measure how frequently academics are publishing articles about any given writer (Porter 2018, 4). Similarly, José Calvo Tello has used the number of pages devoted to an author in a standard literary history as an indicator of prestige when creating the *Corpus of Novels of the Spanish Silver Age* (CONSSA, see Calvo Tello 2021). Ted Underwood and Jordan Sellers have focused instead on reviews in periodicals by creating a dual reference corpus, existing of two samples of both poetry and fiction from the time period 1820-1919 from differing sources (Underwood and Sellers 2015). The first of these was drawn from fourteen trans-Atlantic magazines that selectively reviewed literary works, the

other was taken from a randomized sample from the HathiTrust Digital Library which features a substantially wider array of poetry and fiction.

For quantitative studies of prestige, literary prizes also present corpus building potential by producing discrete shortlists or longlists. James F. English has investigated the relationship between the literary awards industry and the production of cultural value (English 2008). Several corpora have been compiled, such as the Dutch 190 literary prizes corpus (Boudewijn 2020, 33) and the nominees corpus of 50 fiction novels (Koolen 2018, 271), to take stock of the female literary position in the Netherlands as reflected by prize nominations, jury composition and reports, and winnings.

Another strategy to operationalize canonicity through tangible indicators is the use of data from publishing history. In the case of the *European Literary Text Collection* (ELTeC), for example, the criterion of reprint count during a specific, recent period of time (1979-2009) was used to classify novels into two categories: those with two or more reprints, to be understood as still being in active circulation (and hence, canonised), and those with no or just one reprint, understood to be largely marginalized in the contemporary period (Schöch et al. 2021).

Next to this, popular valuation has recently become the focus for corpus building. As part of the aforementioned experiment, the Stanford Literary Lab used rating frequency on GoodReads, since 2007 the world's largest website in recording, sharing, and recommending books, as a selection criterion to cross-reference with expert opinion data (Porter 2018, 3). And the research project *The Riddle of Literary Quality* selected its corpus of 401 novels on the basis of sales and public library loaning figures in the Netherlands between 2009 and 2012, with a focus on selecting those novels with the widest circulation and readership (Koolen et al. 2020).

Pierre Bourdieu's work on the sources of literary distinction, moreover, suggests that beyond the experts and the public, information about co-referencing and peer estimation by artists can act as a proxy for prestige (Bourdieu 1996, 50–51). To the best knowledge of the authors, this avenue for operationalizing literary prestige remains understudied, although it would be an interesting and worth-while challenge to take up.

26.3 Limitations

A first, very fundamental limitation of corpus design practices in the context of canonicity and prestige is that many corpora do not take this into account as a relevant factor at all, often for pragmatic reasons of metadata availability. As a consequence, indicators relevant to the status of the texts with respect to their canonicity or prestige are not included in the metadata, despite the fact that differences in this respect are to be expected. This is the case, in particular, for very large, minimally-curated corpora such as the *Gutenberg Project*. As a consequence, and in particular when small to medium-sized corpora have been designed using the opportunistic model, the effects of availability (especially in digital form) on canonicity are strengthened even further, because there is no conscious intention to counter-balance them, leading to a strong positive bias for canonised works that is not being acknowledged.[1]

Several of the aforementioned studies are limited in scope and representativeness, e.g. they foreground limited data associated with prestige to represent the concept as a whole. From

[1] In the absence of the relevant metadata to check for these effects, it is difficult to provide a clear-cut example for this case. One collections where such an effect could be expected is the *450 Multilingual Novels* collection (Piper and Portelance 2016). More generally, pionieering providers of digital literary texts clearly started out with an unquestioned bias on canonised authors, but in many cases broadened their scope over time, e.g. the *Théâtre classique* corpus (Fièvre 2007).

several viewpoints, this is an understandable strategy to take to corpus building: trading comprehensiveness for feasibility, for example, or negating the problem of compromising data by mapping heterogeneous data sets onto one another. As shown in the previous section, mixed and integrated approaches to corpus building have been successfully undertaken. Yet in a relatively young field in both digital data and methods, comprehensive, multi-dimensional models of literary prestige remain lacking and therefore, no long-term accounts of canon formation and transformation based on substantial sets of data exist. Lastly, it should be noted that current approaches are culturally specific and often focused on Western systems and markets of cultural consumption.

Another potential limitation of current corpus building practices with regard to canonicity or prestige is that they are for the most part focused building a corpus of the canon, or containing only texts belonging to the canon, based on a particular indicator of canonicity. Most of the indicators observed so far, however, could equally well be used to precisely build a corpus that contains texts that are part of the canon (by some measure) as well as texts that are not (by the same measure). Some examples of this practice include the *European Literary Text Collection* mentioned above, which makes it a point to include both canonised and largely forgotten texts, in a bid precisely to allow for comparison of texts from these two groups to be performed. Similarly, Jodie Archer and Matthew Jockers, in their study on textual properties that correlate positively with bestselling novels (Archer and Jockers 2016), included both commercially successful and commercially unsuccessful novels in their corpus. Taking this even further, following the lead of Verboord above, and assuming suitable quantitative rather than categorical indicators, corpora could be built that contain texts of different, defined degrees of canonicity.

References

See works cited and further readings on Zotero.

Citation suggestion

Lisanne van Rossum (2023): "Corpus Building for Canonicity". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/corpus-canon.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{vanrossum_2023_canonicity,
  title = {Corpus Building for Canonicity},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {van Rossum, Lisanne},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/corpus-canon.html},
  langid = {english}
}
```

27 Annotation for Canonicity

Lisanne van Rossum (Amsterdam)

27.1 Annotating the canon

Annotation indicating canonicity or prestige is closely intertwined with the development of selection criteria for a corpus as is shown for example by Algee-Hewitt and McGurl (2015). An annotation scheme for both corpus selection and canonicity requires informed, consistent, and schematic decision-making to maintain internal logic. With this in mind, annotating increasingly complex socio-literary concepts such as canonicity seems a daunting task. The following examples from applied practice present first steps, and are often based on a taxonomy, or classification scheme, of literary value.

27.2 Current practices

In studying the canon, several approaches have been explored concerning annotation for canonicity or prestige. Using categorical principal component analysis (Princals), Marc Verboord has developed an annotation scheme for the prestige status of authors. Princals is a statistical method to reduce variables to a smaller number of components by dividing it up in relative categories, such as "small", "medium" and "large" publishing houses based on their contribution of books published onto the literary market over a given time period (2003, 272). With a data base created from six source types (literary encyclopedias, popular encyclopedias, publishing house status, academic publications, popular prizes, and literary prizes), Verboord created the Institutional Literary Prestige (ILP) system. The ILP system is a classification system covering 502 authors with varying degrees of institutional prestige that juxtaposes popular and literary appeal (2003).

In turn, based on Heydebrand and Winko's 2008 model of literary valuation (see von Heydebrand and Winko 2008; Herrmann, Jacobs, and Piper 2021), Thomas Messerli and Berenike Herrmann have used lemmatization combined with metaphor detection software and manual annotation to investigate how everyday German readers conceptualize literary quality and the reading experience in thematic terms of the thematic categories 'food' (Herrmann and Messerli 2020) and 'motion' (Herrmann and Messerli 2019) in 1.3 million German language online book reviews.

Alternatively, the *Riddle of Literary Quality* project has approached prestige annotation by circumventing a predefined definition of literariness in its entirety and by foregrounding bottom-up perceptions of literary quality from the reading public, who in a National Reader Survey awarded works a Likert scale score between 1 and 7 for literariness (Koolen et al. 2020). A preliminary outcome of the project was that genre and literariness perceptions appeared to be closely related (van Dalen-Oskam 2021; Van Dalen-Oskam 2023). A genre annotation scheme was developed based on the Dutch Uniform Classification Coding system. At the time of the

project, the system was widely applied by publication houses to aid booksellers with the in-store arrangement of books for optimal consumer orientation (Koolen 2018).

Another approach to reader annotation is the Ben-Gurion University of the Negev (BGU) Literary Lab's distant public reading initiative, a large-scale study of the Hebrew novel since its emergence in 1853 to the present day. The collective reading project garnered a total of 525 online questionnaires, filled out by 229 readers about 386 novels. The survey was designed to include narratological properties, such as 'theme,' 'plotting,' and 'structure', and bibliographical properties, such as 'author gender' and 'reception'. The survey also invited the reader to reflect on the perceived readability of the book and the reading experience itself (Dekel and Marienberg-Milikowsky 2021). In practice, no two completed questionnaires of the same novel were identical, a finding that lays bare the complexities of inter-annotator agreement (2021, 246).

27.3 Limitations

It is expected that many more approaches to annotation for canonicity or prestige will be explored, each depending on slightly different research questions or diverging social or historical contexts of the corpus studied. We are still far from a unified and validated annotation standard for canonicity – if that would be possible at all.

References

See works cited and further reading for this chapter on Zotero.

Citation suggestion

Lisanne van Rossum (2023): "Annotation for Canonicity". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/annotation-canon.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{vanrossum_2023_annotationcanonicity,
  title = {Annotation for Canonicity},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {van Rossum, Lisanne},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/annotation-canon.html},
  langid = {english}
}
```

28 Analysis of Canonicity

Lisanne van Rossum (Amsterdam)

28.1 Introduction: Deconstructing the canon

Computational analysis of canonicity and prestige using a large corpus of texts is still rarely done; a sociological approach has been explored more often. The following overview of applied practice will present a sample of analytical techniques currently in use and provide brief explanations of main analytical principles.

28.2 Current applied practice

Much academic enquiry in the area of analysis for canonicity is focused on identifying the stylistic and bibliographical differences between prestigious and non-prestigious contemporary fiction. Piper and Portelance (2016), for example, compared a collection of prizewinning and best-selling fiction (different 'social value groups'), using Linguistic Inquiry and Word Count Software (LIWC), a technique that uses pre-validated word categories to index different textual properties. The researchers also mapped the differences in their corpus across 'genre groups' and constructed a machine learning model to effectively predict whether a newly introduced text was a prize nominee or a Romance novel. Earlier, Kao and Jurafsky (2012) had used LIWC, among others, to develop a computational framework to measure poetic beauty sensations in amateur versus prizewinning poetry reference corpora. Their work also features a useful overview of statistical poetic aesthetics until 2012.

Similarly, Jannidis, Konle, and Leinen (2019) performed an extensive comparison of 'high' and 'low' literary genres published in Germany between 2009 and 2017. A document term matrix with the 8000 most common nouns in the text was used to investigate the stylistic homogeneity with genres, as well as a statistical measure termed *Cosine Delta* of the 2000 most common words in the text. Jannidis, Konle, and Leinen (2019) also employed type-token-ratio and word length to measure stylistic complexity, as well as topic modeling and Zeta-based methods to explore themes and topics in the texts. Topic modeling is a machine learning technique that automatically detects recurring words or phrases in a text or set of texts. Zeta is a keyness measure introduced by John Burrows that is particularly useful to identify content words that are characteristic of one group of texts when compared to another.

van Zundert et al. (2020) investigated the notion of timelessness through stylometric analysis. The researchers compared 'evergreens,' or fiction that remains popular across multiple decades, with former bestseller fiction using TF-IDF vectorization and UMAP dimension reduction. In Natural Language Processing, vectorization is a machine learning technique that trains a model by extracting features from textual data. Dimension reduction then condenses a data set with many features to those features that best express the distance between the data. A popular example of a dimension reduction technique is the Principal Component Analysis.

Also focused on literary value and history, Underwood (2019a) analysed fiction from the mid-nineteenth century to the mid-twentieth century period. Embedded in an argument for 'distant reading' as a lens of interpretation rather than simply a means to an end, Underwood used predictive modelling to evaluate whether a novel or poetry collection's semantic contents were indicators of a higher probability of inclusion in reviews in literature periodicals of the time, and as such elevating its prestige status.

In a similar vein, van Cranenburgh (2016) created a predictive model of textual markers of literary prestige, such as textual complexity, trained on material from novels that were rated in terms of literary quality by Dutch readers. Interestingly, the model identified several works with a discrepancy between style and rating, making reader bias visible. Ashok, Feng, and Choi (2013), too, focused on textual complexity and its correlation with literariness, including factors such as readability indices and lexical choices. The researchers also used part-of-speech (POS) distribution across genres to investigate which word categories were predictive of literary success, and then compared their data set to journalistic writing styles and applied the same analysis to film scripts.

Finally, for the Stanford Literary Lab, J. D. Porter conducted a large-scale investigation of canonicity (2018) by plotting GoodReads reviews (indicative of 'popular appeal') against MLA reviews (indicative of 'elite appeal') of 1406 publishing authors over roughly the last century. These plots, in which authors were listed individually, were condensed down and categorized to map genre spaces, and to further investigate which works clustered together. Porter also constructed a figure that visualized the 'consecration trajectory' in ranking of the 20 most frequently referenced MLA authors throughout the decades 1940-2010.

28.3 Limitations

The current approaches to degrees of canonicity, prestige, or literariness diverge widely and many more options need to be explored in more detail. Future work is expected to zoom in on such questions as which indicators are most important and may be useful for larger scale comparisons into geographical areas, time periods, and perhaps even different kinds of audiences? Which approaches can only be used for individual or local case studies, and which may be useful to analyze longitudinal developments or canonicity issues across language and country borders? How can the historical and sociological context be usefully included in the analysis? Another important element relates to the textual level in relation to languages: if linguistic features are found that may correlate with canonicity or prestige, are these features comparable to those found in corpora in other languages or other time periods etc.? And how should these differences be explored in more detail, in search or more knowledge about canon formation, from both a historical and a (socio-)linguistic perspective? These are only some of the topics that are expected to be addressed in future work.

References

See works cited and further reading for this chapter on Zotero.

Citation suggestion

Lisanne van Rossum (2023): "Analysis of Canonicity". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited

by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/analysis-canon.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{vanrossum_2023_analysiscanonicity,
  title = {Analysis of Canonicity},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {van Rossum, Lisanne},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/analysis-canon.html},
  langid = {english}
}
```

29 Evaluation for Canonicity

Lisanne van Rossum (Amsterdam)

As of yet, no standard method or single tool to study canonicity has been developed, let alone validated and evaluated. To create a computational model that approximates literary value attribution, the current array of analytical techniques is far from standardized (see chapter "Analysis for Canonicity", Chapter 28).

Some scholars offer methodological considerations in the study of literary style, such as Joris van Zundert and Karina van Dalen-Oskam's argument for a holistic, step-by-step approach that integrates qualitative and quantitative research, and for a more expansive curriculum in Humanities training (van Zundert and van Dalen-Oskam 2019).

As Simmons, Nelson, and Simonsohn (2011) demonstrate, conclusions drawn from research can vary greatly depending on those interpreting the results. Much work is still to be done to create robust standards for the evaluation of computational literary studies research into canonicity, and to equip researchers with the skills that can lay a solid foundation for future work.

References

See works cited and further reading for this chapter on Zotero.

Citation suggestion

Lisanne van Rossum (2023): "Evaluation for Canonicity". In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/evaluation-canon.html, DOI: 10.5281/zenodo.7892112.

```
@incollection{vanrossum_2023_evaluationcanonicity,
  title = {Evaluation for Canonicity},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {van Rossum, Lisanne},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/evaluation-canon.html},
  langid = {english}
}
```

Post-Scriptum

Christof Schöch (Trier)

Introduction

With the preparation, conceptualization, writing and publication phases behind us now, in April 2023, we as editors and coordinators have felt a need (and seen potential usefulness to others) for a brief look back on the challenges and benefits of the intense collaborative process of creating this *Survey of Methodological Concerns in Computational Literary Studies*.

We would like to structure this reflection into two concerns, despite the fact that they are of course closely related: the collaborative writing experience as a process of developing and realizing a shared understanding of the survey's goals, structure and format, on the one hand; and the collaborative experiment in writing technology that we conducted at the same time, on the other hand.

Collaborative Writing, or: Towards a shared vision

Although they were originally conceived as five separate and independent surveys, we quickly realized that in order to make the most out of the idea of creating a Survey of Methods in CLS research, a different approach would be more useful. Namely, to think of the survey as a whole as a grid structured by two axes: the key areas of research we wanted to address (such as authorship attribution or canonicity) and the key steps in the research process (such as corpus building or data analysis).

This decision had a number of consequences. For instance, in the grid structure, each area of research, and each step in the research process, would receive an introduction that would be valid and useful to the survey as a whole, thus reducing redundancy and increasing readability. Also, instead of five surveys of medium length, we would end up with a set of 30 short sections forming one coherent whole.

However, many challenges in the implementation of this idea remained. First of all, even with a wiki that allows, in principle, writing and revising right in the online editor, this is clearly not the preferred way of writing for most people, who prefer to use their own writing environments and only upload finalized texts into the wiki in the end. This, however, created challenges for the coherence of the approach and style between sections. In addition, in order to minimize overlap between sections, a clear and shared understanding of the scope of each research area and step in the research process was essential, and needed to be developed in virtual meetings by the geographically-distributed team. Also, with the general and conceptual introductory chapters regarding each research area and step in the research process came a need to agree on a sensible balance between introductory, easily readable textbook prose providing an overview, on the one hand, and more specifically survey-like, detailed documentation of recent practice in CLS research on the other hand.

Connected to this question is of course the question of coverage: with over 1600 publications in the domain of CLS documented in our Zotero database for the period 2011-2020 alone, not every publication could be cited and described in the survey. Still, around 400 publications ended up being mentioned once or multiple times in the survey.

An Experiment in Collaborative Writing Technology

The conceptual work went hand in hand with an open-minded and experimental approach to the question of writing and publication tools. Again, this came with advantages and challenges. Our setup, which may look simple and logical after the fact, but which it took time to settle on, can be described as follows:

Writing with a wiki: Quite early on, we decided that we wanted to use a writing tool that would clearly support a collaborative writing experience, with the possibility of continuous mutual reading and feedback, in order to make sure our shared vision for the survey would become visible in a unified approach to the texts in each section. Therefore, we opted to use a wiki for writing, which thankfully was directly available in our existing Gitlab infrastructure. This is a wiki that is Markdown-based, which turned out to be a good match with our publication framework (see below). It also has a friendly visual editor for a smooth writing experience.

References in Zotero: It was also quite obvious that we wanted to take advantage of the bibliography and corpus of CLS research that we had already assembled for an earlier deliverable (D3.1: Baseline Methodological User Needs Analysis, 2022), in order to give a large and well-document empirical basis to our survey. The corresponding CLS Bibliography was already publicly available on Zotero and was therefore easy to re-use and expand for the survey.

Document transformation using Quarto: An initially less obvious choice was the publication environment. With the grid structure came the desire to publish the survey in a format that would encourage non-linear reading strategies, and we felt that a browser-based reading experience would have clear advantages here over a PDF, which we see more as a derivative, static format for offline reading. Therefore, we wanted to be able to easily create both PDF and HTML versions of our text. The high density and importance of references that a survey inevitably brings with it also meant we needed a robust integration of bibliographic data into the workflow. The solution we found for these requirements was Quarto, an elegant and flexible single-source publication environment. Here, our texts in Markdown from the wiki can be combined with several other files to generate formatted, interactive texts, using the Quarto 'book' format.[1]

Publication on Zenodo and clsinfra.io: Finally, we decided to do the publication of our outputs in two places. On Zenodo, for long-term archiving of all materials, from the Markdown source

[1] A few more details for the technically-minded: Quarto is built on top of the document transformation tool pandoc and can be used in conjunction with plugins for editors such as Visual Studio Code. It relies on a YAML file (`_quarto.yml`) to provide a number of parameters for the book, including metadata, the structure of the book (parts and chapters), and layout parameters for the various output formats (e.g. template file or page format for PDF; theme or base font size for HTML). The bibliographic data is exported from Zotero as a BibTeX file (`references-cited.bib`) and rendered by Quarto in a citation style defined in a CSL file (`chicago-author-date.csl`). Some more display parameters can be defined using a CSS file (`custom.scss`), including things like the banner background or the relative font sizes for headings. For the HTML version, we also embed additional metadata using the `meta` tags in the header (there is helpful information on this on the Zotero website ("Exposing your metadata") and on David I. Verrelli's site ("Metadata tags for academic publications"). A Python script (`metadata_embed.py`) is used to generate and embed the section-level metadata into the `<head>` of each rendered HTML file, based on a pre-populated template file (`metadata_template.md`) and a TSV file (`metadata_source.tsv`) that contains section-level metadata. The PDF file, in turn, is prefixed with a CLS INFRA coversheet for consistency of our deliverables. All of these files are included in the Zenodo upload for documentation.

files to the PDF produced by Quarto, but including the metadata, configuration, BibTeX and HTML files as well. And on the CLS INFRA project website, where the set of static HTML files can easily be placed and provide an intuitive, interactive reading experience.[2]

While we can generally report that the combination of Markdown in a wiki, BibTeX from a Zotero collection, and Quarto for bringing it all together is a great combination, there were of course also technical challenges.

For example, we needed to develop a tagging system to help us track our own work and allow us to generate per-section bibliographies for use by our readers. Because of the combinatorial nature of our grid, this posed challenges for efficiency and precision. Also, at some point our Zotero library would not sync correctly anymore, presumably because of a mismatch between Zotero versions installed on the various computers in our group, making the database instances incompatible with each other. We had to reduce the group of collaborators to the bare minimum and create a new library in order to resolve this issue; clearly not a good solution.

Another example is the way the Markdown files from the Gitlab wiki come together in Quarto. Despite the fact that each wiki on Gitlab is also a repository and can be cloned, making the Markdown files readily available in bulk, it was necessary to move them to a separate repository for production, if only because the file naming conventions in the wiki (based on the title of each page) did not correspond to the way we wanted to design the URLs in the final publication. This is an area where more experiments with a closer integration would probably have been beneficial, because creating duplicates of final versions of section files inevitably, despite our best efforts, led to occasional parallel revisions on both versions of some of the files.

Conclusion

In conclusion, it seems to us that, beyond the publicly available result of this collaborative effort, and despite (or perhaps even because) of the many challenges we encountered, whether regarding the concept, the technical setup, or the teamwork aspects of the project, the most important outcome is probably the learning experience of the entire team. Performing such a collaborative writing process together from the initial idea to the finished product taught as innumerable things. Next time we write a book with a group of twelve people, we will be able to anticipate challenges and pitfalls earlier and be able to create an even smoother experience.

Would we undertake such a collaborative writing project again with a shared understanding of the conceptual structure emerging only in the process? Maybe. Would we do it again using a Gitlab Wiki, a Zotero library and Quarto? Absolutely!

Citation suggestion

Christof Schöch (2023): “Post-Scriptum”. In: *Survey of Methods in Computational Literary Studies* (= D 3.2: Series of Five Short Survey Papers on Methodological Issues). Edited by Christof Schöch, Julia Dudar, Evegniia Fileva. Trier: CLS INFRA. URL: https://methods.clsinfra.io/postscriptum.html, DOI: 10.5281/zenodo.7892112.

[2] Gitlab pages, directly from the CLS INFRA Gitlab instance, would be an alternative and very simple publication channel for the HTML version, given that there are ready-to-use template files for static HTML available in Gitlab.

```
@incollection{schoch_2023_general,
  title = {Post-Scriptum},
  booktitle = {Survey of {{Methods}} in {{Computational
  Literary Studies}} (= {{D}} 3.2: {{Series}} of {{Five Short
  Survey Papers}} on {{Methodological Issues}})},
  author = {Schöch, Christof},
  editor = {Schöch, Christof and Dudar, Julia and Fileva, Evgeniia},
  date = {2023},
  publisher = {{CLS INFRA}},
  location = {{Trier}},
  doi = {10.5281/zenodo.7892112},
  url = {https://methods.clsinfra.io/postscriptum.html},
  langid = {english}
}
```

References

Acerbi, Alberto, Alex Mesoudi, and Marco Smolla. 2022. *Individual-Based Models of Cultural Evolution. A Step-by-Step Guide Using R*. London: Routledge.

Ai, Zhou, Zhang Yijia, Wei Hao, and Lu Mingyu. 2021. "LDA-Transformer Model in Chinese Poetry Authorship Attribution." In *Information Retrieval*, edited by Hongfei Lin, Min Zhang, and Liang Pang, 13026:59–73. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-88189-4_5.

Algee-Hewitt, Mark. 2017. "Distributed Character: Quantitative Models of the English Stage, 1550–1900." *New Literary History* 48 (4): 751–82. https://doi.org/10.1353/nlh.2017.0038.

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Wasler. 2016. "Canon/Archive: Large-scale Dynamics in the Literary Field." *Pamphlets of the Stanford Literary Lab*, no. 11. https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Algee-Hewitt, Mark, and Mark McGurl. 2015. "Between Canon and Corpus: Six Perspectives on 20th-Century Novels." *Pamphlets of the Stanford Literary Lab*, no. 8. https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Andresen, Melanie, Markus Gärtner, Sibylle Hermann, Janina Jacke, Nora Ketschik, Felicitas Lea Kleinkopf, Jonas Kuhn, and Axel Pichler. 2022. "Vorzüge von Auszügen – Urheberrechtlich geschützte Texte in den digitalen Geisteswissenschaften (nach-)nutzen." https://doi.org/10.17175/2022_007.

Andrews, Tara, and Caroline Macé. 2012. "Trees of Texts – Models and Methods for an Updated Theory of Medieval Text Stemmatology." In *Digital Humanities Conference*. https://dh-abstracts.library.virginia.edu/works/1521.

Anthony, Laurence. 2022. "What Can Corpus Software Do?" In *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keeffe and Michael McCarthy, Second. Routledge Handbooks in Applied Linguistics. Abingdon, Oxon ; New York, NY: Routledge.

Antonia, Alexis, Hugh Craig, and Jack Elliott. 2014. "Language Chunking, Data Sparseness, and the Value of a Long Marker List: Explorations with Word n-Grams and Authorial Attribution." *Literary and Linguistic Computing* 29 (2): 147–63. https://doi.org/10.1093/llc/fqt028.

Archer, Jodie, and Matt Jockers. 2016. *The Bestseller Code: Anatomy of the Blockbuster Novel*. New York: St. Martin's Press.

Argamon, Shlomo. 2008. "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations." *Literary and Linguistic Computing* 23 (2): 131–47. https://doi.org/10.1093/llc/fqn003.

Artstein, Ron, and Massimo Poesio. 2008. "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34 (4): 555–96. https://doi.org/10.1162/coli.07-034-R2.

Ashok, Vikas Ganjigunte, Song Feng, and Yejin Choi. 2013. "Success with Style: Using Writing Style to Predict the Success of Novels." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1753–64. Seattle, Washington, USA. https://aclanthology.org/D13-1181.pdf.

Baker, Paul. 2006. *Glossary of Corpus Linguistics*. Edinburgh University Press. https://doi.org/10.1515/9780748626908.

Barbado, Alberto, Víctor Fresno, Ángeles Manjarrés Riesco, and Salvador Ros. 2022. "DISCOPAL: Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels." *Language*

Resources and Evaluation, 1–42. https://link.springer.com/article/10.1007/s10579-021-09557-1.

Barber, Ros. 2018. "Marlowe and Overreaching: A Misuse of Stylometry." *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqy040.

Barbrook, Adrian C., Christopher J. Howe, Norman Blake, and Peter Robinson. 1998. "The Phylogeny of The Canterbury Tales." *Nature* 394 (6696): 839–39. https://doi.org/10.1038/29667.

Barlas, Georgios, and Efstathios Stamatatos. 2020. "Cross-Domain Authorship Attribution Using Pre-trained Language Models." In *Artificial Intelligence Applications and Innovations*, edited by Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, 255–66. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-49161-1_22.

Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. "Individuals, Institutions, and Innovation in the Debates of the French Revolution." *Proceedings of the National Academy of Sciences* 115 (18): 4607–12. https://doi.org/10.1073/pnas.1717729115.

Bergel, Giles, Christopher J Howe, and Heather Windram. 2015. "Lines of Succession in an English Ballad Tradition: The Publishing History and Textual Descent of The Wandering Jew's Chronicle." *Digital Scholarship in the Humanities* 31 (3). https://doi.org/10.1093/llc/fqv003.

Bernstein, Neil, Kyle Gervais, and Wei Lin. 2015. "Comparative Rates of Text Reuse in Classical Latin Hexameter Poetry." *Digital Humanities Quarterly* 9 (3). http://www.digitalhumanities.org/dhq/vol/9/3/000237/000237.html.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024.

———. 1999. *Longman Grammar of Spoken and Written English*. Vol. 2. London: Pearson Education Limited.

———. 2005. "Corpus-Based and Corpus-Driven Analyses of Language Variation and Use." In *The Oxford Handbook of Linguistic Analysis*, Second. Oxford Academic. https://doi.org/10.1093/oxfordhb/9780199677078.013.0008.

Biber, Douglas, Ulla Connor, and Thomas Upton. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure (Studies in Corpus Linguistics)*. Second. Vol. 28. John Benjamins Publishing Co.

Binongo, Jose Nilo G. 2003. "Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution." *Chance* 16 (2): 9–17. https://doi.org/10.1080/09332480.2003.10554843.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. 1st ed. Beijing ; Cambridge [Mass.]: O'Reilly.

Blei, David. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, David, Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, no. 3: 993–1022. https://doi.org/10.1145/2133806.2133826.

Bode, Katherine. 2012. *Reading by Numbers: Recalibrating the Literary Field*. Anthem Press. https://library.oapen.org/handle/20.500.12657/25994.

———. 2020. "Why You Can't Model Away Bias." *Modern Language Quarterly* 81 (1): 95–124. https://doi.org/10.1215/00267929-7933102.

Bollmann, Marcel. 2018. "Normalization of Historical Texts with Neural Network Models." PhD thesis, Bochum: Ruhr-Universität Bochum. https://doi.org/10.13154/294-6213.

Borgo, Mary Elizabeth. 2017. "Toward Sustainable Growth: Lessons Learned Through the Victorian Women Writers Project." *Digital Studies / Le Champ Numérique* 7 (1). https://doi.org/10.16995/dscn.276.

Boudewijn, Petra. 2020. "'And the Award Goes to…' Women on the Dutch Literary Award

Scene." *Journal of Dutch Literature* 11 (1). https://www.journalofdutchliterature.org/index.php/jdl/article/view/198.

Bourdieu, Pierre. 1996. "The Field of Cultural Production: Essays on Art and Literature." *The Journal of Aesthetics and Art Criticism* 54 (1): 88. https://doi.org/10.2307/431688.

Braudel, Fernand, and Immanuel Wallerstein. 2009. "History and the Social Sciences: The Longue Durée." *Review (Fernand Braudel Center)* 32 (2): 171–203. https://www.jstor.org/stable/40647704.

Broadwell, Peter M, and Timothy R Tangherlini. 2017. "Confusing the Modern Breakthrough: Naïve Bayes Classification of Authors and Works." In *DHN*. https://humanit.hb.se/article/view/579.

Brunner, Annelen. 2012. "Automatic Recognition of Speech, Thought and Writing Representation in German Narrative Texts." In *Digital Humanities Conference (DH2012) Book of Abstracts*. Hamburg: ADHO. https://dh-abstracts.library.virginia.edu/works/1415.

———. 2013. "Automatic Recognition of Speech, Thought, and Writing Representation in German Narrative Texts." *Literary and Linguistic Computing* 28 (4): 563–75. https://doi.org/10.1093/llc/fqt024.

Brunner, Annelen, Fotis Jannidis, Tanja Tu, and Lukas Weimer. 2020. "Redewiedergabe in Heftromanen und Hochliteratur." In *Jahreskonferenz 2020 des Verbands Digital Humanities im deutschsprachigen Raum*. Paderborn: DHd-Verband. https://doi.org/10.5281/zenodo.4621813.

Büchler, Marco, Philip R. Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. "Towards a Historical Text Re-use Detection." In *Text Mining: From Ontology Learning to Automated Text Processing Applications*, edited by Chris Biemann and Alexander Mehler, 221–38. Theory and Applications of Natural Language Processing. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-12655-5_11.

Burnard, Lou. 2014. *What Is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources*. Encyclopédie Numérique. Marseille: OpenEditionPress.

Burnard, Lou, Carolin Odebrecht, and Christof Schöch. 2021. "In Search of Comity: TEI for Distant Reading." *Journal of the Text Encoding Initiative*. https://doi.org/10.5281/zenodo.3552488 (preprint).

Burrows, John. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22 (1): 27–47. https://doi.org/10.1093/llc/fqi067.

Butler, Judith. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Thinking Gender. New York: Routledge.

Cafiero, Florian, and Jean-Baptiste Camps. 2019. "Why Molière Most Likely Did Write His Plays." *Science Advances* 5 (11): eaax5489. https://doi.org/10.1126/sciadv.aax5489.

Calvo Tello, José. 2017. "What Does Delta See Inside the Author?: Evaluating Stylometric Clusters with Literary Metadata." In *Humanidades Digitales Hispánicas*. https://cligs.hypotheses.org/aktivitaeten/vortraege.

———. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press. https://doi.org/10.1515/9783839459256.

Camps, Jean-Baptiste, Thibault Clérice, and Ariane Pinche. 2020. "Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis." arXiv. https://doi.org/10.48550/arXiv.2012.03845.

Camps, Jean-Baptiste, and Julien Randon-Furling. 2022. "Lost Manuscripts and Extinct Texts: A Dynamic Model of Cultural Transmission." In *Proceedings of the Computational Humanities Research Conference 2022, CHR 2022, Antwerp, Belgium, December 12-14, 2022*, 3290:198–214. CEUR Workshop Proceedings. CEUR-WS.org. http://ceur-ws.org/Vol-3290/long/_paper3261.pdf.

Carvalho, Loula, and Queiroz. 2016. "Poetry in Prose: Automatic Identification of Verses in Brazilian Literature." In *Digital Humanities Conference 2016 (DH2016) Book of Abstracts*.

https://dh-abstracts.library.virginia.edu/works/2508.
Cavalli-Sforza, Luca L., Paolo Menozzi, and Alberto Piazza. 1994. *The History and Geography of Human Genes*. Text is Free of Markings edition. Princeton, N.J: Princeton University Press.
Chang, Kent K., and Simon DeDeo. 2020. "Divergence and the Complexity of Difference in Text and Culture." *Journal of Cultural Analytics*. https://doi.org/10.22148/001c.17585.
Chaudhuri, Pradim, Tathagata Dasgupta, P. Joseph Dexter, and Krithika Iyer. 2018. "A Small Set of Stylometric Features Differentiates Latin Prose and Verse." *Digital Scholarship in the Humanities* 34 (4): 716–29. https://doi.org/10.1093/llc/fqy070.
Chin, Bertha. 2018. "It's About Who You Know: Social Capital, Hierarchies and Fandom." In *A Companion to Media Fandom and Fan Studies*, edited by Paul Booth, 243–55. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119237211.ch15.
Cinkova, Silvie, and Jan Rybicki. 2020. "Stylometry in a Bilingual Setup." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 977–84. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.123.
Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher Forstall, Roelant Ossewaarde, and Sarah Jacobson. 2012. "The Tesserae Project: Intertextual Analysis of Latin Poetry." *Literary and Linguistic Computing* 28 (2). https://doi.org/10.1093/llc/fqs033.
Coker, Cait, and Kate Ozment. 2019. "Building the Women in Book History Bibliography, or Digital Enumerative Bibliography as Preservation of Feminist Labor." *Digital Humanities Quarterly* 13 (3). http://digitalhumanities.org:8081/dhq/vol/13/3/000428/000428.html.
Coll Ardanuy, Mariona, and Caroline Sporleder. 2014. "Structure-Based Clustering of Novels." In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, 31–39. Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0905.
Collard, Mark, Stephen J. Shennan, and Jamshid J. Tehrani. 2006. "Branching, Blending, and the Evolution of Cultural Similarities and Differences Among Human Populations." *Evolution and Human Behavior* 27: 169–84. https://doi.org/10.1016/j.evolhumbehav.2005.07.003.
Craig, Hugh, and Arthur F. Kinney, eds. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press. https://doi.org/10.1017/CBO9780511605437.
Craig, Hugh, and R. Whipp. 2010. "Old Spellings, New Methods: Automated Procedures for Indeterminate Linguistic Data." *Literary and Linguistic Computing* 25 (1): 37–52. https://doi.org/10.1093/llc/fqp033.
Cutting, Douglass R., David R. Karger, Jan O. Pedersen, and John W. Tukey. 2017. "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections." In *ACM SIGIR Forum*, 51:148–59. New York, NY, USA: ACM. https://doi.org/10.1145/3130348.3130362.
da Silva, Sara Graça, and Jamshid J. Tehrani. 2016. "Comparative Phylogenetic Analyses Uncover the Ancient Roots of Indo-European Folktales." *Royal Society Open Science* 3 (1): 150645. https://doi.org/10.1098/rsos.150645.
Dash, Niladri Sekhar. 2021. *Language Corpora Annotation and Processing*. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-2960-0.
De la Fuente, Eduardo. 2007. "The 'New Sociology of Art': Putting Art Back into Social Science Approaches to the Arts." *Cultural Sociology* 1 (3): 409–25. https://doi.org/10.1177/1749975507084601.
de Paepe, Timothy. 2014. "Visualizing Theatrical Heritage: Computer Modelling as a Tool for Researching the Theatre History of the Low Countries." In *Digital Humanities Conference 2014: Book of Abstracts*. Lausanne: ADHO. https://dh-abstracts.library.virginia.edu/works/2146.
Dekel, Yael, and Itay Marienberg-Milikowsky. 2021. "From Distant to Public Reading The (Hebrew) Novel in the Eyes of Many: The (Hebrew) Novel in the Eyes of Many." *Magazén*, no. 2 (December): JournalArticle_5053. https://doi.org/10.30687/mag/2724-3923/2021/

04/003.

Do Dinh, Erik-Lân, Hannah Wieland, and Iryna Gurevych. 2018. "Weeding Out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1412–24. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1171.

Du, Keli, Julia Dudar, Cora Rok, and Christof Schöch. 2021. "Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness." *Proceedings Computational Humanities Research 2021* 1613: 0073. http://ceur-ws.org/Vol-2989/.

Du, Keli, Julia Dudar, and Christof Schöch. 2022. "Evaluation of Measures of Distinctiveness: Classification of Literary Texts on the Basis of Distinctive Words." https://doi.org/10.48694/JCLS.102.

Durham, William H. 1990. "Advances in Evolutionary Culture Theory." *Annual Review of Anthropology* 19 (1): 187–210. https://doi.org/10.1146/annurev.an.19.100190.001155.

Eder, Maciej. 2011. "Style-Markers in Authorship Attribution : A Cross-Language Study of the Authorial Fingerprint." *Studies in Polish Linguistics* 6 (1). https://ruj.uj.edu.pl/xmlui/handle/item/68325.

———. 2012. "Mind Your Corpus: Systematic Errors in Authorship Attribution." In *Digital Humanities Conference*, 28:4. https://doi.org/10.1093/llc/fqt039.

———. 2013a. "Bootstrapping Delta: A Safety Net in Open-Set Authorship Attribution." In *Digital Humanities Conference*. https://www.semanticscholar.org/paper/Bootstrapping-Delta%3A-a-safety-net-in-open-set-Eder/8fc37de7e6a7cf200d65a12506b5ebf0bda77c4d.

———. 2013b. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." *Literary and Linguistic Computing* 30 (1): 167–82. https://doi.org/10.1093/llc/fqt066.

———. 2013c. "Mind Your Corpus: Systematic Errors in Authorship Attribution." *Literary and Linguistic Computing* 28 (4): 603–14. https://doi.org/10.1093/llc/fqt039.

———. 2015. "Taking Stylometry to the Limits: Benchmark Study on 5,281 Texts from Patrologia Latina." In *Digital Humanities Conference*. Sydney. https://dh-abstracts.library.virginia.edu/works/2364.

———. 2017. "Short Samples in Authorship Attribution: A New Approach." In *Digital Humanities Conference 2017: Book of Abstracts*. https://dh2017.adho.org/abstracts/341/341.pdf.

———. 2022. "Boosting Word Frequencies in Authorship Attribution." arXiv. https://doi.org/10.48550/arXiv.2211.01289.

Eder, Maciej, and Rafał L. Górski. 2023. "Stylistic Fingerprints, POS-tags, and Inflected Languages: A Case Study in Polish." *Journal of Quantitative Linguistics* 30 (1): 86–103. https://doi.org/10.1080/09296174.2022.2122751.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 8 (1): 107. https://doi.org/10.32614/RJ-2016-007.

Ellegård, Alvar. 1962. *A Statistical Method for Determining Authorship: The Junius Letters, 1769-1772*. Gothenburg: University of Gothenburg.

Ellis, Paul D. 2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761676.

English, James F. 2008. *The Economy of Prestige: Prizes, Awards, and the Circulation of Cultural Value*. Cambridge and London: Harvard University Press.

Erlin, Matt. 2017. "Topic Modeling, Epistemology, and the English and German Novel." *Journal of Cultural Analytics*. https://doi.org/ttps://doi.org/10.22148/16.014.

Evert, Stefan. 2008. "Inside the IMS Corpus Workbench." Presentation at the {{IULA}}. Universitat Pompeu Fabra, Barcelona, Spain. https://cwb.sourceforge.io/files/Evert2008_InsideCWB.pdf.

———. 2009. "Corpora and Collocations." In *An International Handbook*, 2:1212–48. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.1212.

Evert, Stefan, and Andrew Hardie. 2011. "Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium." In *Corpus Linguistics 2011*. GBR: University of Birmingham. https://eprints.lancs.ac.uk/id/eprint/62721/.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and Explaining Delta Measures for Authorship Attribution." *Digital Scholarship in the Humanities* 32 (suppl_2). https://doi.org/10.1093/llc/fqx023.

Evert, Stephanie. 2022. "Corpus Encoding and Management Manual CWB Version 3.5." https://cwb.sourceforge.io/files/CWB_Encoding_Tutorial.pdf.

Falk, Michael Gregory. 2015. "Modelling Genre Using Character Networks: The National Tales and Domestic Novels of Maria Edgeworth." In *Digital Humanities 2015: Book of Abstracts*. Sydney: ADHO. https://dh-abstracts.library.virginia.edu/works/2308.

Feldman, Ronen, and James Sanger. 2006. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Feng, Wei Vanessa, and Graeme Hirst. 2013. "Patterns of Local Discourse Coherence as a Feature for Authorship Attribution." *Literary and Linguistic Computing* 29 (2): 191–98. https://doi.org/10.1093/llc/fqt021.

Fièvre, Paul, ed. 2007. *Théâtre classique*. Paris: Université Paris IV-Sorbonne. https://theatre-classique.fr/.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christoph Kittel, Carsten Milling, and Peer Trilcke. 2019. "Programmable Corpora – Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor." In *Jahreskonferenz 2019 des Verbands Digital Humanities im deutschsprachigen Raum*. Mainz und Frankfurt: DHd-Verband. https://doi.org/10.5281/zenodo.4622061.

Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement Among Many Raters." *Psychological Bulletin* 76 (5): 378–82. https://doi.org/10.1037/h0031619.

Forlini, Stefania, Uta Hinrichs, and Bridget Moynihan. 2016. "The Stuff of Science Fiction: An Experiment in Literary History." *Digital Humanities Quarterly* 10 (1). http://www.digitalhumanities.org/dhq/vol/10/1/000228/000228.html.

Forsyth, Richard, and David Holmes. 2018. "The Writeprints of Man: A Stylometric Study of Lafayette's Hand in Paine's 'Rights of Man'." *Digital Humanities Quarterly* 12 (1). http://www.digitalhumanities.org/dhq/vol/12/1/000371/000371.html.

Francis, W. N., and H. Kucera. 1979. "Brown Corpus Manual." Providence: Brown University. http://korpus.uib.no/icame/brown/bcm.html.

Francisco, Virginia, Raquel Hervás, Federico Peinado, and Pablo Gervás. 2012. "EmoTales: Creating a Corpus of Folk Tales with Emotional Annotations." *Language Resources and Evaluation* 46. https://link.springer.com/article/10.1007/s10579-011-9140-5.

Franzini, Greta, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk, and Jan Rybicki. 2018. "Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm." *Frontiers in Digital Humanities* 5. https://www.frontiersin.org/articles/10.3389/fdigh.2018.00004.

Gabrielatos, Costas. 2018. "Keyness Analysis: Nature, Metrics and Techniques." In *Corpus Approaches to Discourse*, edited by Charlotte Taylor and Anna Marchi, 1st Edition, 225–58. Routledge. https://core.ac.uk/download/pdf/227092349.pdf.

Garside, Roger, and Paul Rayson. 1987. "The CLAWS Word-Tagging System." *The Computational Analysis of English: A Corpus-Based Approach*, 30–41. https://www.researchgate.net/publication/2618590_The_CLAWS_Web_Tagger/citations.

Gasparov, Mikhail. 1996. *A History of European Versification*. Oxford, New York: Oxford University Press.

Gavin, Michael. 2014. "Agent-Based Modeling and Historical Simulation." *Digital Humanities Quarterly* 008 (4).

———. 2023. *Literary Mathematics: Quantitative Theory for Textual Studies*. Stanford Text

Technologies. Stanford, California: Stanford University Press.
Gawley, James O, and A Caitlin Diddams. 2017. "Comparing the Intertextuality of Multiple Authors Using Tesserae: A New Technique for Normalization." *Digital Scholarship in the Humanities* 32 (suppl_2): ii53–59. https://doi.org/10.1093/llc/fqx038.
Gehrke, Stefanie. 2016. "Biblissima - Following Medieval Manuscripts and Incunabula Through Their Existence via a Semantic Web Application." In *Digital Humanities 2016: Book of Abstracts*. Krakow: ADHO. https://dh-abstracts.library.virginia.edu/works/2691.
Gemma, Marissa, Frédéric Glorieux, and Jean-Gabriel Frédéric. 2015. "Operationalizing the Colloquial Style: Repetition in 19th-Century American Fiction." *Digital Scholarship in the Humanities* 32 (2). https://doi.org/10.1093/llc/fqv066.
Geyken, Alexander, and Thomas Gloning. 2015. "A Living Text Archive of 15th-19th-Century German. Corpus Strategies, Technology, Organization." In *Historical Corpora: Challenges and Perspectives*, edited by Jost Gippert and Ralf Gehrke. Narr Francke Attempto Verlag.
Gladwin, Alexander A., Matthew J. Lavin, and Daniel M. Look. 2015. "Stylometry and Collaborative Authorship: Eddy, Lovecraft, and 'The Loved Dead'." *Digital Scholarship in the Humanities* 32 (1): 123–40. abc_10.
Gómez-Adorno, Helena, Juan-Pablo Posadas-Duran, Germán Ríos-Toledo, Grigori Sidorov, Gerardo Sierra, Helena Gómez-Adorno, Juan-Pablo Posadas-Duran, Germán Ríos-Toledo, Grigori Sidorov, and Gerardo Sierra. 2018. "Stylometry-Based Approach for Detecting Writing Style Changes in Literary Texts." *Computación y Sistemas* 22 (1): 47–53. https://doi.org/10.13053/cys-22-1-2882.
Gramsci, Antonio. 1991. *Selections from Cultural Writings*. Edited by David Anthony Forgacs. 1. paperback ed. Cambridge: Harvard Univ. Press.
Grayston, K., and G. Herdan. 1959. "The Authorship of the Pastorals in the Light of Statistical Linguistics." *New Testament Studies* 6 (1): 1–15. https://doi.org/10.1017/S0028688500001284.
Grefenstette, Gregory. 1999. "Tokenization." In *Syntactic Wordclass Tagging*, edited by Nancy Ide, Jean Véronis, and Hans van Halteren, 9:117–33. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-9273-4_9.
Gries, Stefan Th. 2008. "Dispersions and Adjusted Frequencies in Corpora." *International Journal of Corpus Linguistics* 13 (4): 403–37. https://doi.org/10.1075/ijcl.13.4.02gri.
———. 2010. "Useful Statistics for Corpus Linguistics." *A Mosaic of Corpus Linguistics: Selected Approaches* 66: 269–91. https://www.researchgate.net/profile/Stefan-Gries-2/publication/267242111_Useful_statistics_for_corpus_linguistics/links/5466341a0cf2f5eb180168f4/Useful-statistics-for-corpus-linguistics.pdf.
———. 2021. "Analyzing Dispersion." In *A Practical Handbook of Corpus Linguistics*, edited by Magali Paquot and Stefan Th. Gries, 99–118. Berlin & New York: Springer. https://stgries.info/research/2020_STG_Dispersion_PHCL.pdf.
Grieve, Jack. 2007. "Quantitative Authorship Attribution: An Evaluation of Techniques." *Literary and Linguistic Computing* 22 (3): 251–70. https://doi.org/10.1093/llc/fqm020.
———. 2023. "Register Variation Explains Stylometric Authorship Analysis." *Corpus Linguistics and Linguistic Theory* 19 (1): 47–77. https://doi.org/10.1515/cllt-2022-0040.
Grieve, Jack, Isobelle Clarke, Emily Chiang, Hannah Gideon, Annina Heini, Andrea Nini, and Emily Waibel. 2019. "Attributing the Bixby Letter Using n-Gram Tracing." *Digital Scholarship in the Humanities* 34 (3): 493–512. https://doi.org/10.1093/llc/fqy042.
Grzybek, Peter. 2014. "The Emergence of Stylometry: Prolegomena to the History of Term and Concept." *Text Within Text - Culture Within Culture*, 58–75. http://www.peter-grzybek.eu/science/publications/2014/grzybek_2014_stylometry.pdf.
Guillory, John. 1993. *Cultural Capital: The Problem of Literary Canon Formation*. Chicago: University of Chicago Press.
Hadjadj, Hassina, and Halim Sayoud. 2021. "Arabic Authorship Attribution Using Synthetic Minority Over-Sampling Technique and Principal Components Analysis for Imbalanced Doc-

uments:" *International Journal of Cognitive Informatics and Natural Intelligence* 15 (4): 1–17. https://doi.org/10.4018/IJCINI.20211001.oa33.

Halvani, Oren, Christian Winter, and Lukas Graner. 2019. "Assessing the Applicability of Authorship Verification Methods." *arXiv:1906.10551 [Cs, Stat]*, June. https://doi.org/10.1145/3339252.3340508.

Hammond, Michael. 2013. "Calculating Syllable Count Automatically from Fixed-Meter Poetry in English and Welsh *." *Literary and Linguistic Computing* 29 (2): 218–33. https://doi.org/10.1093/llc/fqt019.

Han, Jiawei, and Micheline Kamber. 2012. *Data Mining: Concepts and Techniques*. Third. Burlington, MA: Elsevier. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

Harris, Wendell V. 1991. "Canonicity." *PMLA/Publications of the Modern Language Association of America* 106 (1): 110–21. https://doi.org/10.2307/462827.

Hearst, M.arti A., and Jan O. Pedersen. 1996. "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results." In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 76–84. https://doi.org/10.1145/243199.243216.

Heiden, Serge, and Alexei Lavrentiev. 2013. "TXM Platform for Analysis of TEI Encoded Textual Sources." In *Digital Humanities 2013: Book of Abstracts*. Lincoln: ADHO. https://dh-abstracts.library.cmu.edu/works/1713.

Hempfer, Klaus W. 1973. *Gattungstheorie: Information und Synthese*. München: Fink.

———. 2014. "Some Aspects of a Theory of Genre." In *Linguistics and Literary Studies / Linguistik Und Literaturwissenschaft*, edited by Monika Fludernik and Daniel Jacob. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110347500.405.

Henny-Krahmer, Ulrike. 2017. "Bib-ACMé – Bibliografía digital de novelas argentinas, cubanas y mexicanas (1830-1910)." In *Humanidades Digitales Hispánicas (HDH2017): Sociedades, políticas, saberes*. Málaga: HDH. https://hennyu.github.io/hdh_17/index.html#/.

———. 2023. *Genre Analysis and Corpus Design. Nineteenth-Century Spanish-American Novels (1830-1910)*. SIDE 17. Berlin: IDE. https://side17.i-d-e.de/.

Hernández-Lorenzo, Laura, and Joanna Byszuk. 2022. "Challenging Stylometry: The Authorship of the Baroque Play La Segunda Celestina." *Digital Scholarship in the Humanities*, November, fqac063. https://doi.org/10.1093/llc/fqac063.

Herrmann, J. Berenike, Arthur M. Jacobs, and Andrew Piper. 2021. "Computational Stylistics." In *Handbook of Empirical Literary Studies*, edited by Donald Kuiken and Arthur M. Jacobs, 451–86. De Gruyter. https://doi.org/10.1515/9783110645958-018.

Herrmann, J. Berenike, and Thomas Messerli. 2020"... Hungere Schon Nach Dem Nächsten Band. Eine Untersuchung von Metaphern Für Leseerfahrungen in Web 2.0 Literaturrezensionen." https://doi.org/10.5281/zenodo.4621976.

Herrmann, J. Berenike, and Thomas C. Messerli. 2019. "Metaphors We Read by: Finding Metaphorical Conceptualizations of Reading in Web 2.0 Book Reviews." In *Digital Humanities 2020: Book of Abstracts*. ADHO. https://hcommons.org/deposits/item/hc:31569/.

Hettinger, Lena, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho. 2015. "Genre Classification on German Novels." In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, 249–53. Valencia, Spain: IEEE. https://doi.org/10.1109/DEXA.2015.62.

Hettinger, Lena, Fotis Jannidis, Isabella Reger, and Andreas Hotho. 2016. "Significance Testing for the Classification of Literary Subgenres." In *Digital Humanities 2016 Book of Abstracts*. Krakow: ADHO. https://dh-abstracts.library.virginia.edu/works/2630.

Heuser, Ryan, and Long Le-Khac. 2012. "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method." *Standford Literary Lab Pamphlet*,

no. 4. https://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

Hoover, David L. 2015. "Rare N-Grams, Victorian Drama, and Authorship Attribution." In *Digital Humanities Conference 2015 Book of Abstracts*. Sydney: ADHO. https://dh-abstracts.library.virginia.edu/works/2340.

———. 2018. "Authorship Attribution Variables and Victorian Drama: Words, Word-Ngrams, and Character-Ngrams." In *Digital Humanities Conference 2018 Book of Abstracts*. Mexiko City: ADHO. https://dh-abstracts.library.virginia.edu/works/6293.

Hoover, David L. 2007. "Corpus Stylistics, Stylometry, and the Styles of Henry James." *Style* 41 (2): 174–203. https://www.jstor.org/stable/10.5325/style.41.2.174.

Houvardas, John, and Efstathios Stamatatos. 2006. "N-Gram Feature Selection for Authorship Identification." In *Artificial Intelligence: Methodology, Systems, and Applications*, edited by Jérôme Euzenat and John Domingue, 4183:77–86. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11861461_10.

Huber, Eva, and Çağrı Çöltekin. 2020. "Reproduction and Replication: A Case Study with Automatic Essay Scoring." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 5603–13. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.688.

Hughes, James M., Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore. 2012. "Quantitative Patterns of Stylistic Influence in the Evolution of Literature." *Proceedings of the National Academy of Sciences* 109 (20): 7682–86. https://doi.org/10.1073/pnas.1115407109.

Ide, Nancy, and Keith Suderman. 2014. "The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging." *Language Resources and Evaluation* 48 (3): 395–418. https://doi.org/10.1007/s10579-014-9268-1.

Iwata, Yoshimi. 2012. "Tracing the History of Noh Texts by Mathematical Methods Validitating the Application of Phylogenetic Methods to Noh Texts." In *Digital Humanities Conference*. https://researchr.org/publication/dihu-2012.

Jackson, MacDonald P. 2010. "Parallels and Poetry: Shakespeare, Kyd, and "Arden of Faversham." *Medieval & Renaissance Drama in England* 23: 17–33. https://www.jstor.org/stable/24322553.

Jacobs, Arthur M., and Annette Kinder. 2020. "Quasi Error-free Text Classification and Authorship Recognition in a Large Corpus of English Literature Based on a Novel Feature Set." *arXiv:2010.10801 [Cs]*, October. http://arxiv.org/abs/2010.10801.

Jakubíček, Miloš, Adam Kilgarriff, Diana McCarthy, and Pavel Rychlý. 2010. "Fast Syntactic Searching in Very Large Corpora for Many Languages." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 741–47. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. https://aclanthology.org/Y10-1086.

Janicki, Maciej, Kati Kallio, and Mari Sarv. 2022. "Exploring Finnic Written Oral Folk Poetry Through String Similarity." *Digital Scholarship in the Humanities*, July, fqac034. https://doi.org/10.1093/llc/fqac034.

Jannidis, Fotis, Leonard Konle, and Peter Leinen. 2019. "Makroanalytische Untersuchung von Heftromanen." In *Digital Humanities im deutschsprachigen Raum*. https://zenodo.org/record/4622094/files/181_final-KONLE_Leonard_Makroanalytische_Untersuchung_von_Heftromanen.pdf.

Jannidis, Fotis, Leonard Konle, Albin Zehe, Andreas Hotho, and Markus Krug. 2018. "Analysing Direct Speech in German Novels." In *Jahreskonferenz Des DHd-Verbands 2018: Book of Abstracts*. Köln: DHd-Verband. https://doi.org/10.5281/zenodo.4622454.

Jannidis, Fotis, and Gerhard Lauer. 2014. "Burrows's Delta and Its Use in German Literary History." In *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 29–54. Rochester: Camden House. gerhardlauer.de/index.php/download_file/view/335/1/.

Jannidis, Fotis, Isabella Reger, Albin Zehe, Martin Becker, Lena Hettinger, and Andreas Hotho.

2017. "Analyzing Features for the Detection of Happy Endings in German Novels." In *Jahreskonferenz 2017 Des Verbands Digital Humanities Im Deutschsprachigen Raum*. Bern: DHd-Verband. https://doi.org/10.5281/zenodo.4622782.

Jockers, Matt. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Jockers, Matt, and Gabi Kirilloff. 2017. "Understanding Gender and Character Agency in the 19th Century Novel." *Journal of Cultural Analytics*. https://culturalanalytics.org/article/11066-understanding-gender-and-character-agency-in-the-19th-century-novel.

Jockers, Matt, and Daniela Witten. 2010. "A Comparative Study of Machine Learning Methods for Authorship Attribution." *Literary and Linguistic Computing* 25 (2): 215–23. https://doi.org/10.1093/llc/fqq001.

Jungmannová, Lenka, and Petr Plecháč. 2022. "Unsigned Play by Milan Kundera? An Authorship Attribution Study." https://doi.org/10.48550/ARXIV.2212.09879.

Juola, Patrick. 2003. "The Time Course of Language Change." *Computers and the Humanities* 37 (1): 77–96. https://doi.org/10.1023/A:1021839220474.

———. 2015. "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions." *Digital Scholarship in the Humanities* 30 (suppl_1): i100–113. https://doi.org/10.1093/llc/fqv040.

Jurafsky, Daniel, and James.H. Martin. 2023. *Speech and Language Processing*. Draft (3d edition). https://web.stanford.edu/~jurafsky/slp3/.

Jurish, Bryan. 2010. "Comparing Canonicalizations of Historical German Text." In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, 72–77. Uppsala, Sweden: Association for Computational Linguistics. https://aclanthology.org/W10-2209.

Kao, Justine, and Dan Jurafsky. 2012. "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry." In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 8–17. Montréal, Canada: Association for Computational Linguistics. https://aclanthology.org/W12-2502.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schutze. 1997. "Automatic Detection of Text Genre." In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 32–38. Madrid: ACL. https://doi.org/10.3115/976909.979622.

Kestemont, Mike. 2012. "Evaluating Unmasking for Cross-Genre Authorship Verification." In *Digital Humanities 2012: Book of Abstracts*. Hamburg: ADHO. https://dh-abstracts.library.virginia.edu/works/1452.

Kestemont, Mike, and Folgert Karsdorp. 2020. "Estimating the Loss of Medieval Literature with anUnseen Species Model from Ecodiversity." *Computational Humanities Research*. https://ceur-ws.org/Vol-2723/short10.pdf.

Kestemont, Mike, Sara Moens, and Jeroen Deploige. 2013. "Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux." *Literary and Linguistic Computing* 30 (2): 199–224. https://doi.org/10.1093/llc/fqt063.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016a. "Authenticating the Writings of Julius Caesar." *Expert Systems with Applications* 63: 86–96. https://doi.org/10.1016/j.eswa.2016.06.029.

———. 2016b. "Authorship Verification with the Ruzicka Metric." In *Digital Humanities Conference 2016 (DH2016) Book of Abstracts*. Krakow: ADHO. https://dh-abstracts.library.virginia.edu/works/2542.

Kestemont, Mike, Els Stronks, Martine de Bruin, and Tim de Winkel. 2017. *Van Wie Is Het Wilhelmus? De Auteur van Het Nederlandse Volkslied Met de Computer Onderzocht*. Amsterdam: Amsterdam University Press.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: Ten Years On." *Lexicography* 1 (1):

7–36. https://doi.org/10.1007/s40607-014-0009-9.
Kim, Evgeny, Sebastian Pado, and Roman Klinger. 2017. "Prototypical Emotion Developments in Adventures, Romances, and Mystery Stories." In *Digital Humanities 2017: Book of Abstracts*. Montréal: ADHO. https://dh-abstracts.library.virginia.edu/works/3891.
Kocher, Mirco, and Jacques Savoy. 2018. "Evaluation of Text Representation Schemes and Distance Measures for Authorship Linking." *Digital Scholarship in the Humanities* 34 (1): 189–207. https://doi.org/10.1093/llc/fqy013.
Koolen, Corina. 2018. "Reading Beyond the Female: The Relationship Between Perception of Author Gender and Literary Quality." PhD thesis, University of Amsterdam.
Koolen, Corina, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. 2020. "Literary Quality in the Eye of the Dutch Reader: The National Reader Survey." *Poetics* 79 (April): 101439. https://doi.org/10.1016/j.poetic.2020.101439.
Koppel, Moshe, and Jonathan Schler. 2004. "Authorship Verification as a One-Class Classification Problem." In *Twenty-First International Conference on Machine Learning - ICML '04*, 62. Banff, Alberta, Canada: ACM Press. https://doi.org/10.1145/1015330.1015448.
Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. "Computational Methods in Authorship Attribution." *Journal of the American Society for Information Science and Technology* 60 (1): 9–26. https://doi.org/10.1002/asi.20961.
Koppel, Moshe, and Yaron Winter. 2014. "Determining If Two Documents Are Written by the Same Author: Determining If Two Documents Are Written by the Same Author." *Journal of the Association for Information Science and Technology* 65 (1): 178–87. https://doi.org/10.1002/asi.22954.
Krautter, Benjamin. 2020. "Ein Schritt zurück: Distinktive Eigenschaften im deutschsprachigen Drama." In *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum*. Paderborn: DHd-Verband. https://doi.org/10.5281/zenodo.4621893.
Kytö, Merja. 2011. "Corpora and Historical Linguistics." *Revista Brasileira de Linguística Aplicada* 11 (2): 417–57. https://doi.org/10.1590/S1984-63982011000200007.
Labbé, Cyril, and Dominique Labbé. 2001. "Inter-Textual Distance and Authorship Attribution Corneille and Molière." *Journal of Quantitative Linguistics* 8 (3): 213–31. https://doi.org/10.1076/jqul.8.3.213.4100.
Labbé, Dominique. 2007. "Experiments on Authorship Attribution by Intertextual Distance in English*." *Journal of Quantitative Linguistics* 14 (1): 33–80. https://doi.org/10.1080/09296170600850601.
Lassche, Alie, Jan Kostkan, and Kristoffer Nielbo. 2022. "Chronicling Crises: Event Detection in Early Modern Chronicles from the Low Countries." In *Proceedings of the Computational Humanities Research Conference 2022*, 215–30. Antwerp: CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3290/short_paper4697.pdf.
Le Goff, Jacques. 2015. *Must We Divide History into Periods?* Translated by M. B. DeBevoise. European Perspectives. New York Chichester: Columbia University Press.
Leech, Geoffrey. 1993. "Corpus Annotation Schemes." *Literary and Linguistic Computing* 8 (4): 275–81. https://doi.org/10.1093/llc/8.4.275.
Levison, M., A. Q. Morton, and W. C. Wake. 1966. "On Certain Statistical Features of the Pauline Epistles." *The Philosophical Journal* 3: 129–48.
Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. 2014. "Significance Testing of Word Frequencies in Corpora." *Digital Scholarship in the Humanities* 31 (2): 374–97. https://doi.org/10.1093/llc/fqu064.
Long, Hoyt J. 2021. *The Values in Numbers: Reading Japanese Literature in a Global Information Age*. New York: Columbia University Press.
López-Escobedo, Fernanda, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, and Julián Solórzano-Soto. 2013. "Analysis of Stylometric Variables in Long and Short Texts." *Procedia - Social and Behavioral Sciences* 95 (October): 604–11. https://doi.org/10.1016/j.

sbspro.2013.10.688.

Manjavacas, Enrique, Folgert Karsdorp, and Mike Kestemont. 2020. "A Statistical Foray into Contextual Aspects of Intertextuality." *Computational Humanities Research*. https://ceur-ws.org/Vol-2723/long28.pdf.

Manjavacas, Enrique, Brian Long, and Mike Kestemont. 2019. "On the Feasibility of Automated Detection of Allusive Text Reuse." In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 104–14. Minneapolis, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2514.

Mann, Joshua L. 2018. "How Technology Means: Texts, History, and Their Associated Technologies." *Digital Humanities Quarterly* 12 (3). http://digitalhumanities.org:8081/dhq/vol/12/3/000398/000398.html.

Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment." *Computational Linguistics* 41 (3): 437–79. https://doi.org/10.1162/COLI_a_00227.

Mazurko, Anton, and Tomasz Walkowiak. 2020. "Computer Based Stylometric Analysis of Texts in Ukrainian Language." In *Artificial Intelligence and Soft Computing*, edited by Leszek Rutkowski, Rafał Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz, and Jacek M. Zurada, 12416:220–30. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-61534-5_20.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge ; New York: Cambridge University Press.

Meyer, Christine. 2021. *Questioning the Canon: Counter-Discourse and the Minority Perspective in Contemporary German Literature*. De Gruyter. https://www.degruyter.com/document/doi/10.1515/9783110674392/html.

Moisl, Hermann. 2011. "Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora." *Journal of Quantitative Linguistics* 18 (1): 23–52. https://doi.org/10.1080/09296174.2011.533588.

Moravec, Michelle, and Kent K. Chang. 2021. "Feminist Bestsellers: A Digital History of 1970s Feminism." *Journal of Cultural Analytics* 6 (2). https://doi.org/10.22148/001c.22333.

Moretti, Franco. 2000. "The Slaughterhouse of Literature." *Modern Language Quarterly* 61 (1): 207–28. https://doi.org/10.1215/00267929-61-1-207.

———. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Moretti, Franco, and Oleg Sobchuk. 2019. "Hidden In Plain Sight." *New Left Review*, no. 118 (August): 86–115. https://newleftreview.org/issues/ii118/articles/franco-moretti-oleg-sobchuk-hidden-in-plain-sight.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58 (302): 275–309. http://www.jstor.org/stable/2283270.

Mukherjee, Ankhi. 2017. "Canonicity." Oxford: Oxford University Press. https://doi.org/10.1093/obo/9780190221911-0054.

Murrieta-Flores, Patricia, Christopher Donaldson, and Ian Gregory. 2016. "GIS and Literary History: Advancing Digital Humanities Research Through the Spatial Analysis of Historical Travel Writing and Topographical Literature." *Digital Humanities Quarterly* 11 (1). http://www.digitalhumanities.org/dhq/vol/11/1/000283/000283.html.

Muzny, Grace, Mark Algee-Hewitt, and Dan Jurafsky. 2017. "Dialogism in the Novel: A Computational Model of the Dialogic Nature of Narration and Quotations." *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx031.

Nagy, Benjamin. 2021. "Metre as a Stylometric Feature in Latin Hexameter Poetry." *Digital Scholarship in the Humanities* 36 (4): 999–1012. https://doi.org/10.1093/llc/fqaa043.

———. 2023. "Some Stylometric Remarks on Ovid's Heroides and the Epistula Sapphus." *Digital Scholarship in the Humanities*, January, fqac098. https://doi.org/10.1093/llc/fqac09

8.
Navarro-Colorado, Borja. 2017. "A Metrical Scansion System for Fixed-Metre Spanish Poetry." *Digital Scholarship in the Humanities* 33 (1): 112–27. https://doi.org/10.1093/llc/fqx009.
Nini, Andrea. 2018. "An Authorship Analysis of the Jack the Ripper Letters." *Digital Scholarship in the Humanities* 33 (3): 621–36. https://doi.org/10.1093/llc/fqx065.
Nockels, Joe, Paul Gooding, Sarah Ames, and Melissa Terras. 2022. "Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research." *Archival Science* 22 (3): 367–92. https://doi.org/10.1007/s10502-022-09397-0.
Noecker Jr, John, and Michael Ryan. 2012. "Distractorless Authorship Verification." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 785–89. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/238_Paper.pdf.
Ochab, Jeremi K., Joanna Byszuk, Steffen Pielström, and Maciej Eder. 2019. "Identifying Similarities in Text Analysis: Hierarchical Clustering (Linkage) Versus Network Clustering (Community Detection)." In *Digital Humanities 2019: Book of Abstracts*. Utrecht: ADHO. https://doi.org/10.34894/DSVVAC.
Olsen, Mark, Russell Horton, and Glenn Roe. 2011. "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections." *Digital Studies / Le Champ Numérique*. 258-1254-3-PB.
Organisciak, Peter, Sayan Bhattacharyya, Loretta Auvil, J. Stephen Downie, and Beth Plale. 2014. "Large-Scale Text Analysis Through the HathiTrust Research Center." In *Digital Humanities Conference*. https://dh-abstracts.library.virginia.edu/works/1979.
Paige, Nicholas D. 2020. *Technologies of the Novel: Quantitative Data and the Evolution of Literary Systems*. New York: Cambridge University Press.
Paquot, Magali, and Yves Bestgen. 2009. "Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction." In *Corpora: Pragmatics and Discourse*, edited by Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. Brill | Rodopi. https://doi.org/10.1163/9789042029101_014.
Partington, Alan. 2009. "Evaluating Evaluation and Some Concluding Thoughts on CADS." In *Corpus-Assisted Discourse Studies on the Iraq Conflict*, First, 277–320. Routledge.
Pettersson, Eva, Beáta Megyesi, and Jörg Tiedemann. 2013. "An SMT Approach to Automatic Annotation of Historical Text." In *Proceedings from the Workshop on Computational Historical Linguistics at NoDaLiDa 2013*. NEALT. https://cl.lingfil.uu.se/~bea/publ//pettersson-megyesi-tiedemann.pdf.
Pianzola, Federico, Alberto Acerbi, and Simone Rebora. 2020. "Cultural Accumulation and Improvement in Online FanFiction." *Computational Humanities Research*. https://ceur-ws.org/Vol-2723/short8.pdf.
Piotrowski, Michael. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-02146-6.
Piper, Andrew. 2017. "Fictionality." *Journal of Cultural Analytics* 2 (2). https://doi.org/10.22148/16.011.
Piper, Andrew, and Eva Portelance. 2016. "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading." *Post45: Peer Reviewed*. https://post45.org/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/.
Plecháč, Petr. 2021. *Versification and Authorship Attribution*. Charles University in Prague, Karolinum Press.
Porter, J. D. 2018. "Prestige/Popularity." *Pamphlets of the Stanford Literary Lab*, no. 17. https://litlab.stanford.edu/LiteraryLabPamphlet17.pdf.
Powell, Daniel James. 2014. "Seeing Dialogue: Network Visualization of Dramatic Texts." In *Digital Humanities Conference 2014: Book of Abstract*. Lausanne: ADHO. https://dh-

abstracts.library.virginia.edu/works/2055.

Rebora, Simone, J. Berenike Herrmann, Gerhard Lauer, and Massimo Salgaro. 2018a. "Robert Musil, a War Journal, and Stylometry: Tackling the Issue of Short Texts in Authorship Attribution." *Digital Scholarship in the Humanities* 34 (3): 582–605. https://doi.org/10.1093/llc/fqy055.

———. 2018b. "Whose Signal Is It Anyway? A Case Study on Musil for Short Texts in Authorship Attribution." In *Digital Humanities 2018: Book of Abstracts*. Mexiko City: ADHO. https://dh-abstracts.library.virginia.edu/works/6307.

Reeve, Jonathan. 2018. "Does Late Style Exist? New Stylometric Approaches to Variation in Single-Author Corpora." In *Digital Humanities Conference 2018 Book of Abstracts*. Mexiko City: ADHO. https://dh-abstracts.library.virginia.edu/works/6498.

———. 2020. "Text-Matcher." *GitHub Repository*. GitHub. https://doi.org/10.5281/zenodo.3937738.

Reinig, Ines, and Ines Rehbein. 2019. "Metaphor Detection for German Poetry." In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 12. https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_37.pdf.

Reppen, Randi. 2022. "Building a Spoken Corpus : What Are the Basics?" In *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keeffe and Michael McCarthy, Second. Routledge Handbooks in Applied Linguistics. Abingdon, Oxon ; New York, NY: Routledge.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. "OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings." *Applied Sciences* 9 (22): 4853. https://doi.org/10.3390/app9224853.

Rißler-Pipka, Nanette. 2016. "Der falsche Quijote? Autorschaftsattribution für spanische Prosa der frühen Neuzeit." In *DHd 2016 Modellierung, Vernetzung, Visualisierung*, 212–17. Leipzig. http://dhd2016.de/boa.pdf.

Roberts-Smith, Jennifer, Shawn Desouza-Coelho, Teresa M. Dobson, Sandra Gabriele, Omar Rodriguez-Arenas, Stan Ruecker, StéFan Sinclair, et al. 2013. "Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET)." *Digital Humanities Quarterly* 7 (3). http://www.digitalhumanities.org/dhq/vol/7/3/000166/000166.html.

Roberts-Smith, Jennifer, Teresa Dobson, Sandra Gabriele, Omar Rodriguez-Arenas, Stan Ruecker, Stéfan Sinclair, Shawn Desouza-Coelho, Alexandra Kovacs, and Daniel So. 2013. "Visualising Theatre Historiography: Judith Thompson's White Biting Dog (1984 and 2011) in the Simulated Environment for Theatre (SET)." *Digital Studies / Le Champ Numérique* 3 (2). https://doi.org/10.16995/dscn.240.

Roe, Glenn, Clovis Gladstone, Mark Olsen, and Robert Morrissey. 2016. "Digging into ECCO: Identifying Commonplaces and Other Forms of Text Reuse at Scale." In *Digital Humanities 2016: Book of Abstracts*. Krakow. https://hal.science/hal-03211786.

Roelli, Philipp, ed. 2020. *Handbook of Stemmatology: History, Methodology, Digital Approaches*. De Gruyter. https://doi.org/10.1515/9783110684384.

Romanowska, Iza, Colin D. Wren, and Stefani A. Crabtree. 2021. *Agent-Based Modeling for Archaeology: Simulating the Complexity of Societies*. Santa Fe: Santa Fe Institute Press.

Ruiz Fabo, Pablo, Clara Martínez Cantón, and José Calvo Tello. 2018. "DISCO: Diachronic Spanish Sonnet Corpus." In *Digital Humanities Im Deutschsprachigen Raum*. http://e-spacio.uned.es/fez/eserv/bibliuned:363-Pruiz2/Ruiz_Fabo_Pablo_DISCO_corpus.pdf.

Rybicki, Jan. 2015. "Vive La Différence : Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies." *Digital Scholarship in the Humanities* 31 (4): 746–61. https://doi.org/10.1093/llc/fqv023.

Rybicki, Jan, and Maciej Eder. 2011. "Deeper Delta Across Genres and Languages: Do We

Really Need the Most Frequent Words?" *Literary and Linguistic Computing* 26 (3): 315–21. https://doi.org/10.1093/llc/fqr031.

Rychlý, Pavel. 2007. "Manatee/Bonito - A Modular Corpus Manager." In *Proceedings of Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University. https://nlp.fi.muni.cz/raslan/2007/papers/12.pdf.

Sack, Graham Alexander. 2013. "Simulating Plot: Towards a Generative Model of Narrative Structure." In *Digital Humanities Conference*. https://dh-abstracts.library.virginia.edu/works/1779.

Sánchez-Martínez, Felipe, Isabel Martínez-Sempere, Xavier Ivars-Ribes, and Rafael C. Carrasco. 2013. "An Open Diachronic Corpus of Historical Spanish: Annotation Criteria and Automatic Modernisation of Spelling." arXiv. https://doi.org/10.48550/arXiv.1306.3692.

Savoy, Jacques. 2020a. "Elena Ferrante: A Case Study in Authorship Attribution." In *Machine Learning Methods for Stylometry*, 191–210. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-53360-1_8.

———. 2020b. *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-53360-1.

Schmidt, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Schneider, Gerold, Eva Pettersson, and Michael Percillier. 2017. "Comparing Rule-based and SMT-based Spelling Normalisation for English Historical Texts." In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 40–46. Gothenburg: Linköping University Electronic Press. https://aclanthology.org/W17-0508.

Schöch, Christof. 2017a. "Aufbau von Datensammlungen." In *Digital Humanities: Eine Einführung*, edited by Fotis Jannidis, Hubertus Kohle, and Malte Rehbein, 223–32. Stuttgart: Metzler.

———. 2017b. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

———. 2018a. "Théâtre Classique, Paul Fièvre Ed." *RIDE – A Review Journal for Digital Editions and Resources*, no. 8. https://ride.i-d-e.de/issues/issue-8/theatre-classique/.

———. 2018b. "Zeta für die kontrastive Analyse literarischer Texte Theorie, Implementierung, Fallstudie." In *Quantitative Ansätze in den Literatur- und Geisteswissenschaften*, edited by Toni Bernhart, Marcus Willand, Sandra Richter, and Andrea Albrecht, 77–94. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110523300-004.

———. 2023. "Repetitive Research: A Conceptual Space and Terminology of Replication, Reproduction, Re-Implementation, Re-Analysis, and Re-Use in Computational Literary Studies." *International Journal of Digital Humanities*, March, preprint. https://doi.org/10.21203/rs.3.rs-2657846/v1.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. "Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen." https://doi.org/10.17175/2020_006.

Schöch, Christof, Tomaž Erjavec, Roxana Patraș, and Diana Santos. 2021. "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives." *Modern Languages Open* 1 (25): 1–19. https://doi.org/10.3828/mlo.v0i0.364.

Schöch, Christof, Maria Hinzmann, Julia Röttgermann, Katharina Dietz, and Anne Klee. 2022. "Smart Modelling for Literary History." *International Journal of Humanities and Arts Computing* 16 (1): 78–93. https://doi.org/10.3366/ijhac.2022.0278.

Schöch, Christof, and Allen Riddell. 2014. "Progress Through Regression. Modeling Style Across Genre in French Classical Theater." In *Digital Humanities Conference 2014: Book of Abstract*. Lausanne: ADHO. https://doi.org/10.5281/zenodo.7785294.

Schöch, Christof, Daniel Schlör, Stefanie Popp, Annelen Brunner, Ulrike Henny, and José Calvo Tello. 2016. "Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels." In *Digital Humanities Conference 2016 (DH2016) Book of Abstracts*. Krakow: ADHO. https://dh-abstracts.library.virginia.edu/works/2515.

Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Daniel Becker, and Andreas Hotho. 2018. "Burrows' Zeta: Exploring and Evaluating Variants and Parameters." In *Digital Humanities 2018: Book of Abstracts*. Mexiko City: ADHO. https://dh-abstracts.library.virginia.edu/works/6327.

Schöch, Christof, Karina Van Dalen-Oskam, Maria Antoniak, Fotis Jannidis, and David Mimno. 2020. "Replication and Computational Literary Studies." In *Digital Humanities Conference 2020: Book of Abstracts*. Ottawa: ADHO. https://dh2020.adho.org/wp-content/uploads/2020/07/178_ReplicationandComputationalLiteraryStudies.html.

Schößler, Franziska. 2010. *Einführung in die Gender Studies*. Berlin: Akademie Verlag.

Schröter, Julian, Keli Du, Julia Dudar, Cora Rok, and Christof Schöch. 2021. "From Keyness to Distinctiveness – Triangulation and Evaluation in Computational Literary Studies." *Journal of Literary Theory* 15 (1-2): 81–108. https://doi.org/10.1515/jlt-2021-2011.

Schumacher, Mareike, and Marie Flüh. 2020. "m*w Figurengender zwischen Stereotypisierung und literarischen und theoretischen Spielräumen Genderstereotype und -bewertungen in der Literatur des 19 Jahrhunderts." In *Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum: Book of Abstracts*. Paderborn: DHd-Verband. https://doi.org/10.5281/zenodo.4621891.

Šeļa, Artjoms, Petr Plecháč, and Alie Lassche. 2022. "Semantics of European Poetry Is Shaped by Conservative Forces: The Relationship Between Poetic Meter and Meaning in Accentual-Syllabic Verse." *PLOS ONE* 17 (4). https://doi.org/10.1371/journal.pone.0266556.

Šeļa, Artjoms, and Oleg Sobchuk. 2017. "The Shortest Species: How the Length of Russian Poetry Changed (1750–1921)." *Studia Metrica Et Poetica* 4 (1): 66–84. https://doi.org/10.12697/smp.2017.4.1.03.

Seo, Jangwon, and W. Bruce Croft. 2008. "Local Text Reuse Detection." In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 571–78. SIGIR '08. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1390334.1390432.

Shang, Wenyi, and Ted Underwood. 2021. "Improving Measures of Text Reuse in English Poetry: 16th International Conference on Diversity, Divergence, Dialogue, iConference 2021." Edited by Katharina Toeppe, Hui Yan, and Samuel Kai Chu. *Diversity, Divergence, Dialogue - 16th International Conference, iConference 2021, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 469–77. https://doi.org/10.1007/978-3-030-71292-1_36.

Sharmaa, Aniruddha, Yuerong Hu, Peizhen Wu, Wenyi Shang, Shubhangi Singhal, and Ted Underwood. 2020. "The Rise and Fall of Genre Differentiation inEnglish-Language Fiction." In *Computational Humanities Conference 2020*. Amsterdam: CEUR. https://ceur-ws.org/Vol-2723/long27.pdf.

Shutova, Ekaterina. 2017. "Annotation of Linguistic and Conceptual Metaphor." In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 1073–1100. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-024-0881-2_40.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Sirkin, R. Mark. 2006. *Statistics for the Social Sciences*. Third. SAGE Publications, Inc.

Smaldino, Paul E. 2017. "Models Are Stupid, and We Need More of Them." In *Computational Social Psychology*, 311–31. Routledge. https://doi.org/10.4324/9781315173726-14.

Smith, David A., Ryan Cordel, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson.

2014. "Detecting and Modeling Local Text Reuse." In *IEEE/ACM Joint Conference on Digital Libraries*, 183–92. https://doi.org/10.1109/JCDL.2014.6970166.

So, Richard Jean, Hoyt Long, and Yuancheng Zhu. 2018. "Race, Writing, and Computation: Racial Difference and the US Novel, 1880-2000." *Journal of Cultural Analytics* 3 (2). https://doi.org/10.22148/16.031.

Sobchuk, Oleg. 2016. "The Evolution of Dialogues: A Quantitative Study of Russian Novels (1830–1900)." *Poetics Today* 37 (1): 137–54. https://doi.org/10.1215/03335372-3452643.

Spencer, Matthew, Elizabeth A Davidson, Adrian C Barbrook, and Christopher J Howe. 2004. "Phylogenetics of Artificial Manuscripts." *Journal of Theoretical Biology* 227 (4): 503–11. https://doi.org/10.1016/j.jtbi.2003.11.022.

Sperberg-McQueen, Michael, and Lou Burnard. 1994. "Guidelines for Electronic Text Encoding and Interchange." Oxford: Text Encoding Initiative. https://tei-c.org/guidelines/p5/.

Stamatatos, Efstathios. 2009. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60 (3): 538–56. https://doi.org/10.1002/asi.21001.

———. 2013. "On the Robustness of Authorship Attribution Based on Character *N*-Gram Features." *Journal of Law and Policy* 21 (2). https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/7.

Stefanowitsch, Anatol. 2020. *Corpus Linguistics: A Guide to the Methodology*. Vol. 7. Language Science Press. https://doi.org/10.5281/ZENODO.3735822.

Storey, Grant, and David Mimno. 2020. "Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus." *Journal of Cultural Analytics* 5 (2). https://doi.org/10.22148/001c.13680.

Sturgeon, Donald. 2017. "Unsupervised Identification of Text Reuse in Early Chinese Literature." *Digital Scholarship in the Humanities* 33 (3). https://doi.org/10.1093/llc/fqx024.

Suzuki, Takafumi, Shuntaro Kawamura, Fuyuki Yoshikane, Kyo Kageura, and Akiko Aizawa. 2012. "Co-Occurrence-Based Indicators for Authorship Analysis." *Literary and Linguistic Computing* 27 (2): 197–214. https://doi.org/10.1093/llc/fqs011.

Tagliamonte, Sali A. 2011. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Malden, MA: Wiley-Blackwell.

Talbot, Mary M. 2019. *Language and Gender*. Third. Cambridge, UK ; Medford, MA: Polity Press.

Tehrani, Jamshid J. 2013. "The Phylogeny of Little Red Riding Hood." *PLOS ONE* 8 (11): e78871. https://doi.org/10.1371/journal.pone.0078871.

Tehrani, Jamshid, Quan Nguyen, and Teemu Roos. 2015. "Oral Fairy Tale or Literary Fake? Investigating the Origins of Little Red Riding Hood Using Phylogenetic Network Analysis." *Digital Scholarship in the Humanities* 31 (3): 611–36. https://doi.org/10.1093/llc/fqv016.

Throne, Jeremy. 2014. "Modeling the Communications Circuit: An Agent-based Approach to Reading in 'N-Dimensions'." In *Complexity and the Human Experience*. Jenny Stanford Publishing.

Thuillard, Marc, Julien d'Huy, Yuri Berezkin, and Jean-Loïc Le Quellec. 2018. "A Large-Scale Study of World Myths." *Trames Journal of the Humanities and Social Sciences* 22 (January): 407–24.

Tollebeek, Jo, Marc Boone, and Karel Van Nieuwenhuyse. 2022. *Een canon van Vlaanderen : motieven en bezwaren*. Vol. 78. KVAB Press. http://hdl.handle.net/1854/LU-8766324.

Traub, Myriam, and Jacco van Ossenbruggen. 2015. "Workshop on Tool Criticism in the Digital Humanities." Amsterdam: CWI. https://ir.cwi.nl/pub/23500.

Trilcke, Peer, Frank Fischer, and Mathias Göbel. 2016. "Dramen als small worlds? Netzwerkdaten zur Geschichte und Typologie deutschsprachiger Dramen 1730-1930." In *Jahreskonferenz 2016 des Verbands Digital Humanities im deutschsprachigen Raum*. Leipzig: DHd-Verband. https://zenodo.org/record/3526423#.ZCcSynZBy3A.

Trilcke, Peer, Frank Fischer, Mathias Göbel, and Dario Kampkaspar. 2016. "Theatre Plays

as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930." In *Digital Humanities Conference 2016: Book of Abstracts*. Krakow: ADHO. https://dh2016.adho.org/abstracts/360.

Trilcke, Peer, Frank Fischer, and Dario Kampkaspar. 2015. "Digital Network Analysis of Dramatic Texts." In *Digital Humanities Conference 2015: Book of Abstracts*. Sydney: ADHO. https://dh-abstracts.library.virginia.edu/works/2228.

Tsvetkov, Yulia, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. "Metaphor Detection with Cross-Lingual Model Transfer." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 248–58. Baltimore, Maryland: ACL. https://doi.org/10.3115/v1/P14-1024.

Tuccinardi, Enrico. 2016. "An Application of a Profile-Based Method for Authorship Verification: Investigating the Authenticity of Pliny the Younger's Letter to Trajan Concerning the Christians." *Digital Scholarship in the Humanities* 32 (2): 435–47. https://doi.org/10.1093/llc/fqw001.

Tuzzi, Arjuna, and Michele A. Cortelazzo, eds. 2018. *Drawing Elena Ferrante's Profile: Workshop Proceedings, Padova, 7 September 2017*. Padova: Padova UP.

Underwood, Ted. 2014. "The Proportional Sizes of Genres in Eighteenth- and Nineteenth-Century English-Language Books." In *Digital Humanities 2014: Book of Abstracts*. Lausanne: ADHO. https://dh-abstracts.library.virginia.edu/works/2109.

———. 2017. "The Life Cycles of Genres." *Journal of Cultural Analytics* 2 (2). https://doi.org/10.22148/16.005.

———. 2019a. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.

———. 2019b. "Do We Understand the Outlines of Literary History?" In *Distant Horizons*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226612973.003.0001.

Underwood, Ted, David Bamman, and Sabrina Lee. 2018. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics* 3 (2). https://doi.org/10.22148/16.019.

Underwood, Ted, Michael L. Black, Loretta Auvil, and Boris Capitanu. 2013. "Mapping Mutable Genres in Structurally Complex Volumes." In *2013 IEEE International Conference on Big Data*, 95–103. Silicon Valley, CA, USA: IEEE. https://doi.org/10.1109/BigData.2013.6691676.

Underwood, Ted, Kevin Kiley, Wenyi Shang, and Stephen Vaisey. 2022. "Cohort Succession Explains Most Change in Literary Culture." *Sociological Science* 9 (May): 184–205. https://doi.org/10.15195/v9.a8.

Underwood, Ted, and Jordan Sellers. 2012. "The Emergence of Literary Diction." *Journal of Digital Humanities* 1 (2). https://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers/.

———. 2015. "How Quickly Do Literary Standards Change?" Figshare. https://doi.org/10.6084/m9.figshare.1418394.v1.

———. 2016. "The Longue Durée of Literary Prestige." *Modern Language Quarterly* 77 (3): 321–44. https://doi.org/10.1215/00267929-3570634.

Underwood, Ted, and Richard Jean So. 2021. "Can We Map Culture?" *Journal of Cultural Analytics* 6 (3). https://doi.org/10.22148/001c.24911.

van Cranenburgh, A. W. 2016. "Rich Statistical Parsing and Literary Language." PhD thesis, Amsterdam: University of Amsterdam. https://hdl.handle.net/11245/1.543163.

van Dalen-Oskam, Karina. 2021. *Het Raadsel Literatuur: Is Literaire Kwaliteit Meetbaar?* Amsterdam: Amsterdam University Press.

Van Dalen-Oskam, Karina. 2023. *The Riddle of Literary Quality: A Computational Approach*. Amsterdam: Amsterdam University Press. https://eadh.org/projects/riddle-literary-quality.

van Dalen-Oskam, Karina, and Joris van Zundert. 2007. "Delta for Middle Dutch—Author and Copyist Distinction in Walewein." *Literary and Linguistic Computing* 22 (3): 345–62. https://doi.org/10.1093/llc/fqm012.

Van Deinsen, Lieke, Anthe Sevenants, and Freek Van de Velde. 2022. "De Nederlandstalige Literaire Canon(s) anno 2022. Een enquête naar de literaire klassieken: rapportage (voorpublicatie)." Gent: KANTL. https://ctb.kantl.be/assets/files/pages/files/De_Nederlandstalige_literaire_canon(s)_anno_2022_-_Een_enqu%C3%AAte_naar_de_literaire_klassieken_Rapport_(voorpublicatie).pdf.

Van Rees, C. J. 1983. "How a Literacy Work Becomes a Masterpiece: On the Threefold Selection Practised by Literary Criticism." *Poetics* 12 (4-5): 397–417. https://doi.org/10.1016/0304-422X(83)90015-3.

———. 1987. "How Reviewers Reach Consensus on the Value of Literary Works." *Poetics* 16 (3-4): 275–94. https://doi.org/10.1016/0304-422X(87)90008-8.

van Rees, Kees, and Jeroen Vermunt. 1996. "Event History Analysis of Authors' Reputation: Effects of Critics' Attention on Debutants' Careers." *Poetics* 23 (5): 317–33. https://doi.org/10.1016/0304-422X(96)00002-2.

Van Rossum, Lisanne M., and Artjoms Šeļa. 2022. "CLS INFRA D4.1 Skills Gap Analysis." In. Krakow: CLS INFRA. https://doi.org/10.5281/ZENODO.6401858.

van Zundert, Joris J., Raymond A. Mar, Karina van Dalen-Oskam, Emily Temple, Isabel Bowman, Farzaneh Heidari, and Ahn Nguyen. 2020. "Features of Timelessness: Intermediate Report on a Quest for Stylistic Features That Mark Literary Canonicity." In *Digital Humanities 2020 Book of Abstracts*. Ottowa, Canada. https://dh-abstracts.library.virginia.edu/works/9675.

van Zundert, Joris J., and Karina van Dalen-Oskam. 2019. "Joris en Karina's Holistisch Letterkundig Onderzoeksbureau." *Tijdschrift voor Nederlandse Taal- en Letterkunde (Journal of Dutch Linguistics and Literature)* 125 (4): 357–72. https://doi.org/10.5117/TNTL2019.4.006.ZUND.

Vega García-Luengos, Germán. 2021. "Las Comedias de Lope de Vega: Confirmaciones de Autoría y Nuevas Atribuciones Desde La Estilometría (I)." *Talía. Revista de Estudios Teatrales* 3 (May): 91–108. https://doi.org/10.5209/tret.74625.

Verboord, Marc. 2003. "Classification of Authors by Literary Prestige." *Poetics* 31 (3-4): 259–81. https://doi.org/10.1016/S0304-422X(03)00037-8.

Vickers, Brian. 2008. "Thomas Kyd, Secret Sharer." *Times Literary Supplement*, no. 13.

———. 2011. "Shakespeare and Authorship Studies in the Twenty-First Century." *Shakespeare Quarterly* 62 (1): 106–42. https://doi.org/10.1353/shq.2011.0004.

———. 2012. "Identifying Shakespeare's Additions to *The Spanish Tragedy* (1602): A New(er) Approach." *Shakespeare* 8 (1): 13–43. https://doi.org/10.1080/17450918.2012.660283.

von Heydebrand, Renate, and Simone Winko. 2008. "13. The Qualities of Literatures: A Concept of Literary Evaluation in Pluralistic Societies." In *Linguistic Approaches to Literature*, edited by Willie van Peer, 4:223–39. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lal.4.16hey.

Wadsworth, Fabian B., Jérémie Vasseur, and David E. Damby. 2016. "Evolution of Vocabulary in the Poetry of Sylvia Plath." *Digital Scholarship in the Humanities* 32 (3). https://doi.org/10.1093/llc/fqw026.

Wang, Haining, Xin Xie, and Allen Riddell. 2021. "Cross-Register Authorship Attribution Using Vernacular and Classical Chinese Texts." In *DH Benelux Conference*. https://doi.org/10.5281/zenodo.4886595.

Weerasinghe, Janith, Rhia Singh, and Rachel Greenstadt. 2021. "Feature Vector Difference Based Authorship Verification for Open-World Settings: 2021 Working Notes of CLEF - Conference and Labs of the Evaluation Forum, CLEF-WN 2021." *CEUR Workshop Proceedings* 2936: 2201–7. https://ceur-ws.org/Vol-2696/paper_125.pdf.

Weidman, Sean G., and James O'Sullivan. 2017. "The Limits of Distinctive Words: Re-

evaluating Literature's Gender Marker Debate." *Digital Scholarship in the Humanities* 33 (2): 374–90. https://doi.org/10.1093/llc/fqx017.

Wernimont, Jacqueline, and Julia Flanders. 2011. "Possible Worlds: Authorial Markup and Digital Scholarship." In *Digital Humanities 2011: Book of Abstracts*. Lincoln: ADHO. https://dh-abstracts.library.virginia.edu/works/1317.

Wilkens, Matthew. 2017. "Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction." *Cultural Analytics* 2 (2). https://doi.org/10.22148/16.009.

Winter, Bodo, and Paul-Christian Bürkner. 2021. "Poisson Regression for Linguists: A Tutorial Introduction to Modelling Count Data with Brms." *Language and Linguistics Compass* 15 (11): e12439. https://doi.org/10.1111/lnc3.12439.

Wolfe, Joanna. 2002. "Annotation Technologies: A Software and Research Review." *Computers and Composition* 19 (4): 471–97. https://doi.org/10.1016/S8755-4615(02)00144-5.

Worsham, Joseph, and Jugal Kalita. 2018. "Genre Identification and the Compositional Effect of Genre in Literature." In *Proceedings of the 27th International Conference on Computational Linguistics*, 1963–73. Santa Fe: ACL. https://aclanthology.org/C18-1167.

Youngblood, Mason, Karim Baraghith, and Patrick E. Savage. 2020. "Phylogenetic Reconstruction of the Cultural Evolution of Electronic Music via Dynamic Community Detection (1975-1999)." *arXiv:2011.02460 [q-Bio, Stat]*, November. http://arxiv.org/abs/2011.02460.

Zeldes, Amir, Anke Lüdeling, Julia Ritz, and Christian Chiarcos. 2009. "ANNIS: A Search Tool for Multi-Layer Annotated Corpora." In *Corpus Linguistics 2009*. Liverpool. https://doi.org/10.18452/13437.